Generalised Regression Models

GRM: Solutions 2 **Semester 1, 2022–2023**

---

1. To show that the distributions are members of the exponential family write the pdf (probability function, discrete case) in the form

$$\exp\{a(y)b(\theta)+c(\theta)+d(y)\},$$

where $a(y)$ and $d(y)$ are functions of y, and $b(\theta)$ and $c(\theta)$ are functions of θ.

To determine the means and the variances use the general results obtained in the lecture notes:

$$\mathrm{E}\{a(Y)\} = -\frac{c'(\theta)}{b'(\theta)} \quad \text{and} \quad \mathrm{var}\{a(Y)\} = \frac{b''(\theta)c'(\theta)-c''(\theta)b'(\theta)}{\{b'(\theta)\}^3}$$

(i) Exponential:

$$f(y;\theta) = \theta e^{-y\theta} = \exp(\log\theta - y\theta)$$

$$a(y)=y \quad b(\theta)=-\theta \quad c(\theta)=\log\theta \quad d(y)=0$$

$$\mathrm{E}(Y)=\mathrm{E}\{a(Y)\} = -\frac{c'(\theta)}{b'(\theta)} = \frac{1}{\theta}$$

$$\mathrm{var}(Y)=\mathrm{var}\{a(Y)\} = \frac{b''(\theta)c'(\theta)-c''(\theta)b'(\theta)}{\{b'(\theta)\}^3} = \frac{(0)\frac{1}{\theta}-\left(-\frac{1}{\theta^2}\right)(-1)}{(-1)^3} = \frac{1}{\theta^2}$$

(ii) Binomial:

$$\begin{aligned} f(y;\theta) &= \binom{m}{y}\theta^y(1-\theta)^{m-y} \\ &= \exp\left\{y(\log\theta-\log(1-\theta))+m\log(1-\theta)+\log\binom{m}{y}\right\} \end{aligned}$$

$$a(y)=y \quad b(\theta)=\log\left(\frac{\theta}{1-\theta}\right) \quad c(\theta)=m\log(1-\theta) \quad d(y)=\log\binom{m}{y}$$

$$\mathrm{E}(Y)=\mathrm{E}\{a(Y)\} = -\frac{c'(\theta)}{b'(\theta)} = -\frac{(-m/(1-\theta))}{\frac{1}{\theta}+\frac{1}{1-\theta}} = m\theta$$

$$\begin{aligned} \mathrm{var}(Y)=\mathrm{var}\{a(Y)\} &= \frac{b''(\theta)c'(\theta)-c''(\theta)b'(\theta)}{\{b'(\theta)\}^3} \\ &= \frac{\left(-\frac{1}{\theta^2}+\frac{1}{(1-\theta)^2}\right)\left(\frac{-m}{(1-\theta)}\right)-\frac{-m}{(1-\theta)^2}\left(\frac{1-\theta+\theta}{\theta(1-\theta)}\right)}{\left(\frac{1}{\theta}+\frac{1}{1-\theta}\right)^3} \\ &= m\theta(1-\theta) \end{aligned}$$

(iii) Gamma:

$$f(y;\theta) = \frac{y^{\phi-1}\theta^{\phi}e^{-y\theta}}{\Gamma(\phi)} = \exp\{(\phi-1)\log y+\phi\log\theta-y\theta-\log\Gamma(\phi)\}$$

$$a(y)=y \quad b(\theta)=-\theta \quad c(\theta)=\phi\log\theta \quad d(y)=(\phi-1)\log y-\log\Gamma(\phi).$$

$$\mathrm{E}(Y)=\mathrm{E}\{a(Y)\} = -\frac{c'(\theta)}{b'(\theta)} = \frac{\phi}{\theta}$$

$$\begin{aligned} \mathrm{var}(Y)=\mathrm{var}\{a(Y)\} &= \frac{b''(\theta)c'(\theta)-c''(\theta)b'(\theta)}{\{b'(\theta)\}^3} \\ &= \frac{(0)\frac{\phi}{\theta}-\left(-\frac{\phi}{\theta^2}\right)(-1)}{-1} = \frac{\phi}{\theta^2} \end{aligned}$$

(iv) Neg. bin.:
$$\begin{aligned} f(y;\theta) &= \binom{y+r-1}{r-1}\theta^r(1-\theta)^y \\ &= \exp\left\{r\log\theta + y\log(1-\theta) + \log\binom{y+r-1}{r-1}\right\} \end{aligned}$$

$$a(y) = y \quad b(\theta) = \log(1-\theta) \quad c(\theta) = r\log\theta \quad d(y) = \log\binom{y+r-1}{r-1}$$

$$\mathrm{E}(Y) = \mathrm{E}\{a(Y)\} = -\frac{c'(\theta)}{b'(\theta)} = \frac{r(1-\theta)}{\theta}$$

$$\begin{aligned} \mathrm{var}(Y) = \mathrm{var}\{a(Y)\} &= \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{\{b'(\theta)\}^3} \\ &= \frac{\left(-\frac{1}{(1-\theta)^2}\right)\left(\frac{r}{\theta}\right) - \left(-\frac{r}{\theta^2}\right)\left(-\frac{1}{(1-\theta)}\right)}{-\frac{1}{(1-\theta)^3}} \\ &= \frac{r(1-\theta)}{\theta^2} \end{aligned}$$

The distributions (i)–(iv) are all in canonical form since $a(y) = y$. Natural parameters

$$\text{(i) } b(\theta) = -\theta \quad \text{(ii) } b(\theta) = \log\left(\frac{\theta}{1-\theta}\right) \quad \text{(iii) } b(\theta) = -\theta \quad \text{(iv) } b(\theta) = \log(1-\theta)$$

2. (a) The maximum likelihood estimates are determined by θ which maximizes the log likelihood

$$\begin{aligned} l(\theta) &= \log\left[\prod_{i=1}^{n}\exp\{a(y_i)b(\theta) + c(\theta) + d(y_i)\}\right] \\ &\equiv b(\theta)\sum_{i=1}^{n}a(y_i) + nc(\theta) \end{aligned}$$

Differentiate the log likelihood to obtain the score function

$$U(\theta) = l'(\theta) = b'(\theta)\sum_{i=1}^{n}a(y_i) + nc'(\theta)$$

Solving $U(\widehat{\theta}) = 0$ determines the maximum likelihood estimate for θ (provided that $\widehat{\theta}$ corresponds to a maximum, i.e. $l''(\widehat{\theta}) < 0$, and $l(\theta)$ is differentiable at $\widehat{\theta}$).

(i) Exp:
$$a(y) = y \quad b'(\theta) = -1 \quad c'(\theta) = \frac{1}{\theta}$$

$$\begin{aligned} & U(\widehat{\theta}) = 0 \\ \Leftrightarrow\ & b'(\widehat{\theta})\sum_{i=1}^{n}a(y_i) + nc'(\widehat{\theta}) = 0 \\ \Leftrightarrow\ & (-1)\sum_{i=1}^{n}y_i + n\frac{1}{\widehat{\theta}} = 0 \\ \Leftrightarrow\ & \text{MLE } \widehat{\theta} = \frac{n}{\sum_{i=1}^{n}y_i} \end{aligned}$$

(ii) Bin (m is known):
$$a(y) = y \quad b'(\theta) = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)} \quad c'(\theta) = -\frac{m}{1-\theta}$$

$$\begin{aligned} & U(\widehat{\theta}) = 0 \\ \Leftrightarrow\ & b'(\widehat{\theta})\sum_{i=1}^{n}a(y_i) + nc'(\widehat{\theta}) = 0 \\ \Leftrightarrow\ & \frac{1}{\widehat{\theta}(1-\widehat{\theta})}\sum_{i=1}^{n}y_i - \frac{nm}{(1-\widehat{\theta})} = 0 \\ \Leftrightarrow\ & \text{MLE } \widehat{\theta} = \frac{\sum_{i=1}^{n}y_i}{nm} \end{aligned}$$

(iii) Gamma:

$$a(y) = y \quad b'(\theta) = -1 \quad c'(\theta) = \frac{\phi}{\theta}$$

$$U(\widehat{\theta}) = 0$$

$$\begin{aligned}
&\Leftrightarrow \quad b'(\widehat{\theta})\sum_{i=1}^{n} a(y_i) + nc'(\widehat{\theta}) = 0 \\
&\Leftrightarrow \quad (-1)\sum_{i=1}^{n} y_i + n\frac{\phi}{\widehat{\theta}} = 0 \\
&\Rightarrow \quad \text{MLE } \widehat{\theta} = \frac{n\phi}{\sum_{i=1}^{n} y_i} = \frac{\phi}{\bar{y}}
\end{aligned}$$

(iv) Neg. bin.:

$$a(y) = y \quad b'(\theta) = -\frac{1}{(1-\theta)} \quad c'(\theta) = \frac{r}{\theta}$$

$$U(\widehat{\theta}) = 0$$

$$\begin{aligned}
&\Leftrightarrow \quad b'(\widehat{\theta})\sum_{i=1}^{n} a(y_i) + nc'(\widehat{\theta}) = 0 \\
&\Leftrightarrow \quad -\frac{1}{1-\widehat{\theta}}\sum_{i=1}^{n} y_i + \frac{nr}{\widehat{\theta}} = 0 \\
&\Leftrightarrow \quad -\widehat{\theta}n\bar{y} + nr(1-\widehat{\theta}) = 0 \\
&\Leftrightarrow \quad (n\bar{y} + nr)\widehat{\theta} = nr \\
&\Leftrightarrow \quad \text{MLE } \widehat{\theta} = \frac{r}{\bar{y}+r}
\end{aligned}$$

(b) The asymptotic (large sample) distribution of the maximum likelihood estimator is

$$\widehat{\theta} \sim N(\theta, I_\theta^{-1}) \quad \text{where} \quad I_\theta = -\mathrm{E}\left(\frac{d^2 l}{d\theta^2}\right) = -\mathrm{E}(U').$$

I_θ is the Fisher information for θ contained in the sample $y_1, \ldots, y_n$.
The information may be obtained by using

$$I_\theta = -\mathrm{E}(U') = -b''(\theta)\sum_{i=1}^{n} \mathrm{E}\{a(Y_i)\} - nc''(\theta) = -b''(\theta)n\mathrm{E}\{Y\} - nc''(\theta)$$

since iid case, and $a(Y) = Y$.
Thus, using E(Y) given in Question 1, we have (asymptotically)

(i) $I_\theta = -n\{(0)\frac{1}{\theta} + (-\frac{1}{\theta^2})\} = \frac{n}{\theta^2} \Rightarrow \mathrm{var}(\widehat{\theta}) = I_\theta^{-1} = \frac{\theta^2}{n}$

(ii) $I_\theta = -n\{(-\frac{1}{\theta^2} + \frac{1}{(1-\theta)^2})(m\theta) + \frac{-m}{(1-\theta)^2}\} = \frac{nm}{\theta(1-\theta)} \Rightarrow \mathrm{var}(\widehat{\theta}) = I_\theta^{-1} = \frac{\theta(1-\theta)}{nm}$

(iii) $I_\theta = -n\{(0)(\frac{\phi}{\theta}) + (-\frac{\phi}{\theta^2})\} = \frac{n\phi}{\theta^2} \Rightarrow \mathrm{var}(\widehat{\theta}) = I_\theta^{-1} = \frac{\theta^2}{n\phi}$

(iv) $I_\theta = -n\{(-\frac{1}{(1-\theta)^2})(\frac{r(1-\theta)}{\theta}) + (-\frac{r}{\theta^2})\} = \frac{nr}{\theta^2(1-\theta)} \Rightarrow \mathrm{var}(\widehat{\theta}) = I_\theta^{-1} = \frac{\theta^2(1-\theta)}{nr}$

3. (a) The ith response Y_i has probability density function

$$(2\pi)^{-\frac{1}{2}}\sigma^{-1}\exp\left\{-\frac{1}{2}\sigma^{-2}(y_i - \beta_0 - \beta_1 x_i)^2\right\} \quad (i = 1, \ldots, n),$$

so the joint probability density function is the product of these,

$$(2\pi)^{-\frac{1}{2}n}\sigma^{-n}\exp\left\{-\frac{1}{2}\sigma^{-2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right\} = (2\pi)^{-\frac{1}{2}n}\sigma^{-n}\exp\left(-\frac{1}{2}\sigma^{-2}Q\right).$$

(b) This is the likelihood function, L, and is maximized with respect to β_0 and β_1 when Q is minimized, so that the maximum likelihood estimates of these two parameters are also the least squares estimates.

(c) The logarithm of the likelihood is (apart from an additive constant)

$$l = \ln L = -n \ln \sigma - \frac{1}{2}\sigma^{-2} Q.$$

The derivative of this with respect to σ is

$$\frac{\partial l}{\partial \sigma} = -n\sigma^{-1} + \sigma^{-3} Q.$$

Equating this to zero, the maximum likelihood estimate of σ^2 is found to be $n^{-1}RSS$, where RSS denotes the residual sum of squares, which is the minimum of Q with respect to β_0 and β_1.

4. Although not asked for, we can easily check the formulae given for $E(Y)$ and $\text{var}(Y)$ using the moment generating function of the normal distribution. If $Z \sim N(0,1)$

$$\begin{aligned}
M_Z(t) &= E\left(e^{tZ}\right) \\
&= \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz \\
&= e^{\frac{t^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(z-t)^2\right) dz \\
&= e^{\frac{t^2}{2}} [1] = e^{\frac{t^2}{2}}
\end{aligned}$$

Hence if $X \sim N(\mu, \sigma^2)$, $X = \mu + \sigma Z$

$$\begin{aligned}
M_X(t) &= E\left(e^{tX}\right) \\
&= e^{t\mu} E\left(e^{t\sigma Z}\right) \\
&= e^{t\mu} M_Z(t\sigma) \\
&= e^{t\mu + \frac{1}{2}\sigma^2 t^2}
\end{aligned}$$

Now, if Y has the log normal distribution, we can write $\log(Y) = X$, or $Y = e^X$. Thus

$$\begin{aligned}
E(Y) &= E(e^X) = M_X(1) = e^{\mu + \frac{1}{2}\sigma^2} \\
E(Y^2) &= E(e^{2X}) = M_X(2) = e^{2\mu + 2\sigma^2} \\
\text{var}(Y) &= E(Y^2) - [E(Y)]^2 = e^{2\mu + 2\sigma^2} - e^{2\mu + \sigma^2} = e^{2\mu + \sigma^2}\left[e^{\sigma^2} - 1\right]
\end{aligned}$$

Thus,

$$\nu = e^{\mu + \frac{\sigma^2}{2}} \quad \text{and} \quad \tau = \left(e^{\sigma^2} - 1\right) e^{2\mu + \sigma^2}$$

(a) For a random sample $x_1, \ldots, x_n$ from $N(\mu, \sigma^2)$ we know that the MLEs are

$$\widehat{\mu} = \bar{x} \qquad \widehat{\sigma}^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

The mapping from (μ, σ) to (ν, τ) is 1–1 (you can check that the determinant of the Jacobian is non-zero), hence the MLEs of ν and τ are just the corresponding function of $\widehat{\mu}$ and $\widehat{\sigma}^2$.

$$\begin{aligned}
\widehat{\nu} &= \exp\left[\widehat{\mu} + \frac{1}{2}\widehat{\sigma}^2\right] \\
\widehat{\tau} &= \exp\left[2\widehat{\mu} + \widehat{\sigma}^2\right]\left(e^{\widehat{\sigma}^2} - 1\right)
\end{aligned}$$

where $\widehat{\mu}$ and $\widehat{\sigma}^2$ are given in terms of $x_1, \ldots, x_n$ as above. Clearly the formulae can be written in terms of the y_i, by replacing x_i by $\log(y_i)$.

(b) If σ^2 is known we simply replace $\widehat{\sigma}^2$ by σ^2 in the equations above, yielding

$$\begin{aligned}\widehat{\nu} &= \exp\left[\widehat{\mu}+\frac{1}{2}\sigma^2\right]\\ \widehat{\tau} &= \exp\left[2\widehat{\mu}+\sigma^2\right]\left(e^{\sigma^2}-1\right)\end{aligned}$$

Now obtain the expectation of $\widehat{\nu}$ for case (b):

$$\widehat{\nu} = \exp\left[\widehat{\mu}+\frac{1}{2}\sigma^2\right] = e^{\frac{1}{2}\sigma^2}e^{\widehat{\mu}}$$

Thus

$$E(\widehat{\nu}) = e^{\frac{1}{2}\sigma^2}E(e^{\widehat{\mu}}) = e^{\frac{1}{2}\sigma^2}E(e^{\bar{X}})$$

Since we know $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, we can use the general form of the moment generating function of the normal distribution to write

$$E(e^{\bar{X}}) = M_{\bar{X}}(1) = e^{\mu+\frac{1}{2}\frac{\sigma^2}{n}}$$

Hence

$$\begin{aligned}E(\widehat{\nu}) &= e^{\mu+\frac{1}{2}\sigma^2+\frac{1}{2}\frac{\sigma^2}{n}}\\ &= \nu e^{\frac{1}{2}\frac{\sigma^2}{n}}\end{aligned}$$

We see that $\widehat{\nu}$ is biased as an estimator of ν, but as $n \to \infty$, bias $\to 0$.

Fisher's information is given by,

$$I(\nu) = \frac{I(\mu)}{\left(\frac{d\nu}{d\mu}\right)^2}$$

Now

$$\begin{aligned}\nu &= e^{\mu+\frac{\sigma^2}{2}}\\ \frac{d\nu}{d\mu} &= e^{\mu+\frac{\sigma^2}{2}} = \nu\end{aligned}$$

We know that for a normal distribution $I(\mu) = \frac{n}{\sigma^2}$. Thus

$$I(\nu) = \frac{\frac{n}{\sigma^2}}{\nu^2} = \frac{n}{\sigma^2\nu^2}$$

and

$$\text{var}(\widehat{\nu}) \to I^{-1}(\nu) = \frac{\sigma^2\nu^2}{n}$$

as $n \to \infty$.

5. (a) $l(\theta;y) = \log f(y;\theta) = y\theta - c(\theta) + d(y)$ and $U(\theta) = \frac{dl}{d\theta} = y - c'(\theta)$
E$(U) = 0$ so E$(Y - c'(\theta)) = 0$. Thus E$(Y) = c'(\theta) = \mu$.

$$\text{var}(U) = -\text{E}\left(\frac{d^2l}{d\theta^2}\right) = -\text{E}(-c''(\theta)) = c''(\theta)$$

$$\text{var}(Y) = \text{var}(U + c'(\theta)) = \text{var}(U) = c''(\theta)$$

(b) $\frac{dl}{d\theta} = y - \mu$ since $\mu = c'(\theta)$

(c) $\frac{dl}{d\mu} = \frac{dl}{d\theta}\frac{d\theta}{d\mu} = \frac{\frac{dl}{d\theta}}{\frac{d\mu}{d\theta}} = \frac{y-\mu}{\text{var}(Y)}$ since $\frac{d\mu}{d\theta} = c''(\theta) = \text{var}(Y)$.

6. (a) $\frac{dl}{d\theta} = \frac{y-c'(\theta)}{a(\phi)}$ and for a single observation $\mathrm{E}(\frac{dl}{d\theta}) = 0$. Thus $\mathrm{E}(Y - c'(\theta)) = 0$ i.e. $\mathrm{E}(Y) = c'(\theta)$. Also

$$\mathrm{var}\left(\frac{dl}{d\theta}\right) = -\mathrm{E}\left(\frac{d^2 l}{d\theta^2}\right) = -\mathrm{E}\left[-\frac{c''(\theta)}{a(\phi)}\right]$$

by the same argument as in the 1-parameter case. Thus

$$\begin{aligned}
\frac{1}{a^2(\phi)}\mathrm{var}(Y) &= \mathrm{var}\left(\frac{dl}{d\theta}\right) \\
&= -\mathrm{E}\left[-\frac{c''(\theta)}{a(\phi)}\right] \\
&= \frac{c''(\theta)}{a(\phi)} \\
\mathrm{var}(Y) &= a(\phi)c''(\theta)
\end{aligned}$$

(b) **Normal:**

$$\begin{aligned}
f(y;\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \\
&= \exp\left(-\frac{y^2-2y\mu+\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right) \\
&= \exp\left(\frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right)
\end{aligned}$$

which is of the form required with $\theta = \mu$ and $\phi = \sigma^2$ and

$$c(\mu) = \frac{\mu^2}{2};\;\; a(\sigma^2) = \sigma^2;\;\; d(y,\sigma^2) = -\frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)$$

Gamma:

$$\begin{aligned}
\log\left(f(y;\lambda,\alpha)\right) &= -\lambda y + \log(\lambda) + (\alpha-1)\log(y\lambda) - \log\left(\Gamma(\alpha)\right) \\
&= -\lambda y + \alpha\log(\lambda) + (\alpha-1)\log(y) - \log\left(\Gamma(\alpha)\right)
\end{aligned}$$

Defining $\theta = -\frac{\lambda}{\alpha}$ and $\phi = \alpha$

$$\begin{aligned}
\log\left(f(y;\theta,\phi)\right) &= \theta\phi y + \phi\log(-\theta\phi) + (\phi-1)\log(y) - \log\left(\Gamma(\phi)\right) \\
&= \theta\phi y + \phi\log(-\theta) + \phi\log(\phi) + (\phi-1)\log(y) - \log\left(\Gamma(\phi)\right) \\
&= \frac{y\theta - (-\log(-\theta))}{\phi^{-1}} + \phi\log(\phi) + (\phi-1)\log(y) - \log\left(\Gamma(\phi)\right)
\end{aligned}$$

which is of the required form, with

$$c(\theta) = -\log(-\theta);\;\; a(\phi) = \phi^{-1};\;\; d(y,\phi) = \phi\log(\phi) + (\phi-1)\log(y) - \log\left(\Gamma(\phi)\right)$$

(c) Let $X = \log(Y)$, $Y = e^X$. The Jacobian of the transformation is

$$\frac{dy}{dx} = e^x = y$$

$$f_Y(y) = \frac{f_X(\log(y))}{\left|\frac{dy}{dx}\right|} = \frac{1}{y}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(\log(y)-\mu)^2}{2\sigma^2}\right)$$

$$\log f_Y(y;\mu,\sigma^2) = -\log(y) - \frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left[(\log(y))^2 - 2\mu(\log(y)) + \mu^2\right]$$

Clearly there is no term in the form $y\theta$ where θ is a function of μ and σ^2, so this density cannot be written in the form required.

7. For a single observation y

$$\begin{aligned} l(\theta;y) &= \log f(y;\theta) \\ \frac{\partial l}{\partial \theta_j} &= \frac{1}{f}\frac{\partial f}{\partial \theta_j} \end{aligned}$$

Then

$$\begin{aligned} E\left(\frac{\partial l}{\partial \theta_j}\right) &= \int \frac{1}{f}\frac{\partial f}{\partial \theta_j} f(y;\theta)dy \\ &= \int \frac{\partial f}{\partial \theta_j} dy \\ &= \frac{\partial}{\partial \theta_j}\int f(y;\theta)dy \\ &= \frac{\partial}{\partial \theta_j} 1 \\ &= 0 \end{aligned}$$

and

$$\begin{aligned} \frac{\partial^2 l}{\partial \theta_j \partial \theta_k} &= -\frac{1}{f^2}\frac{\partial f}{\partial \theta_j}\frac{\partial f}{\partial \theta_k} + \frac{1}{f}\frac{\partial^2 f}{\partial \theta_j \partial \theta_k} \\ &= -\frac{\partial l}{\partial \theta_j}\frac{\partial l}{\partial \theta_k} + \frac{1}{f}\frac{\partial^2 f}{\partial \theta_j \partial \theta_k} \end{aligned}$$

Taking expectations gives

$$\begin{aligned} E\left(\frac{\partial^2 l}{\partial \theta_j \partial \theta_k}\right) &= -E\left(\frac{\partial l}{\partial \theta_j}\frac{\partial l}{\partial \theta_k}\right) + \int \frac{1}{f}\frac{\partial^2 f}{\partial \theta_j \partial \theta_k} f(y;\theta)dy \\ &= -E\left(\frac{\partial l}{\partial \theta_j}\frac{\partial l}{\partial \theta_k}\right) + \frac{\partial^2}{\partial \theta_j \partial \theta_k}\int f(y;\theta)dy \\ &= -E\left(\frac{\partial l}{\partial \theta_j}\frac{\partial l}{\partial \theta_k}\right) + \frac{\partial^2}{\partial \theta_j \partial \theta_k} 1 \\ &= -E\left(\frac{\partial l}{\partial \theta_j}\frac{\partial l}{\partial \theta_k}\right) + 0 \end{aligned}$$

Thus

$$\operatorname{cov}\left(\frac{\partial l}{\partial \theta_j}, \frac{\partial l}{\partial \theta_k}\right) = E\left(\frac{\partial l}{\partial \theta_j}\frac{\partial l}{\partial \theta_k}\right) - 0 = -E\left(\frac{\partial^2 l}{\partial \theta_j \partial \theta_k}\right)$$

To prove the same results for a random sample $y_1, \ldots, y_n$, we can argue in either of two ways:

(a) For $i = 1, \ldots, n$ the random variables $\frac{\partial l(\theta;Y_i)}{\partial \theta_j}$ are iid. Thus

$$E\left(\frac{\partial}{\partial \theta_j}\sum_{i=1}^n l(\theta;Y_i)\right) = \sum_{i=1}^n E\left(\frac{\partial}{\partial \theta_j} l(\theta;Y_i)\right) = 0,$$

and

$$\begin{aligned} \operatorname{cov}\left(\frac{\partial}{\partial \theta_j}\sum_{i=1}^n l(\theta;Y_i), \frac{\partial}{\partial \theta_k}\sum_{i=1}^n l(\theta;Y_i)\right) &= \sum_{i=1}^n \operatorname{cov}\left(\frac{\partial l(\theta;Y_i)}{\partial \theta_j}, \frac{\partial l(\theta;Y_i)}{\partial \theta_k}\right) \\ &= \sum_{i=1}^n -E\left(\frac{\partial^2 l(\theta;Y_i)}{\partial \theta_j \partial \theta_k}\right) \\ &= -E\left(\frac{\partial^2 \sum_{i=1}^n l(\theta;Y_i)}{\partial \theta_j \partial \theta_k}\right) \end{aligned}$$

(b) Let $L(\theta;\mathbf{y}) = \prod_{i=1}^{n} f(y_i;\theta)$ and $l(\theta;\mathbf{y}) = \log(L)$

$$\frac{\partial l}{\partial \theta_j} = \frac{1}{L}\frac{\partial L}{\partial \theta_j}$$

$$\begin{aligned}
E\left(\frac{\partial l}{\partial \theta_j}\right) &= \int \frac{1}{L}\frac{\partial L}{\partial \theta_j} L d\mathbf{y} \\
&= \int \frac{\partial L}{\partial \theta_j} d\mathbf{y} \\
&= \frac{\partial}{\partial \theta_j} \int L d\mathbf{y} \\
&= \frac{\partial}{\partial \theta_j} \int \cdots \int f(y_1;\theta)\cdots f(y_n;\theta) dy_1 \cdots dy_n \\
&= \frac{\partial}{\partial \theta_j} \int f(y_1;\theta) dy_1 \cdots \int f(y_n;\theta) dy_n \\
&= \frac{\partial}{\partial \theta_j} 1 \\
&= 0
\end{aligned}$$

And, as before,

$$\begin{aligned}
\frac{\partial^2 l}{\partial \theta_j \partial \theta_k} &= -\frac{\partial l}{\partial \theta_j}\frac{\partial l}{\partial \theta_k} + \frac{1}{L}\frac{\partial^2 L}{\partial \theta_j \partial \theta_k} \\
E\left(\frac{\partial^2 l}{\partial \theta_j \partial \theta_k}\right) &= -E\left(\frac{\partial l}{\partial \theta_j}\frac{\partial l}{\partial \theta_k}\right) + \int \frac{1}{L}\frac{\partial^2 L}{\partial \theta_j \partial \theta_k} L d\mathbf{y} \\
&= -E\left(\frac{\partial l}{\partial \theta_j}\frac{\partial l}{\partial \theta_k}\right) + \frac{\partial^2}{\partial \theta_j \partial \theta_k} \int L d\mathbf{y} \\
&= -E\left(\frac{\partial l}{\partial \theta_j}\frac{\partial l}{\partial \theta_k}\right) + \frac{\partial^2}{\partial \theta_j \partial \theta_k} \int \cdots \int f(y_1;\theta)\cdots f(y_n;\theta) dy_1 \cdots dy_n \\
&= -E\left(\frac{\partial l}{\partial \theta_j}\frac{\partial l}{\partial \theta_k}\right) + \frac{\partial^2}{\partial \theta_j \partial \theta_k} \int f(y_1;\theta) dy_1 \cdots \int f(y_n;\theta) dy_n \\
&= -E\left(\frac{\partial l}{\partial \theta_j}\frac{\partial l}{\partial \theta_k}\right) + \frac{\partial^2}{\partial \theta_j \partial \theta_k} 1 \\
&= -E\left(\frac{\partial l}{\partial \theta_j}\frac{\partial l}{\partial \theta_k}\right) + 0
\end{aligned}$$

Thus

$$\text{cov}\left(\frac{\partial l}{\partial \theta_j}, \frac{\partial l}{\partial \theta_k}\right) = E\left(\frac{\partial l}{\partial \theta_j}\frac{\partial l}{\partial \theta_k}\right) - 0 = -E\left(\frac{\partial^2 l}{\partial \theta_j \partial \theta_k}\right)$$