Derivatives of the Likelihood of a Single Observation from a Generalised Linear Hidden Markov Model

1 Structure of the Model

The likelihood of a hidden Markov model is calculated (by means of a recursive procedure) from likelihoods of single observations which are of the form $f(y, \boldsymbol{\theta})$. These expressions may be either probability density functions or a probality mass functions. The symbol y represents an observation and $\boldsymbol{\theta}$ represents a vector of parameters upon which the distribution in question depends. These parameters depend in turn on the underlying state of the hidden Markov chain and in general upon other predictors (in addition to "state"). The dependence of $\boldsymbol{\theta}$ upon the predictors will involve further parameters. In order to effect the recursive procedure referred to above, we need to calculate the first and second derivatives, with respect to all of the parameters that are involved, of the single observation likelihoods $f(y, \boldsymbol{\theta})$.

We are concerned with five distributions: Gaussian, Poisson, Binomial, Db ("discretised beta") and Multinom. In the Poisson and Binomial cases the models are generalised linear models. In the Gaussian and Db cases the models are "something like, but not exactly" generalised linear models. In the case of the Multinom (or "discnp" — discrete non-parametric) distribution the model in question bears some relationship to a generalised linear model but is of a substantialy different form. We shall use the expression "extended generalised hidden Markov models". to describe the collection of all models under consideration, including those based on the Gaussian, Db and Multinom distributions.

In the case of the Gaussian distribution $\boldsymbol{\theta} = (\mu, \sigma)^\top$ where μ is the mean and σ is the standard deviation of the distribution. In the cases of the Poisson and Binomial distributions $\boldsymbol{\theta}$ is actually a scalar (which we consequently write simply at θ). For the Poisson distribution θ is equal to λ, the Poisson mean, and for the Binomial distribution θ is equal to p, the binomial success probability. In the case of the Db distribution, $\boldsymbol{\theta}$ is equal to $(\alpha, \beta)^\top$ the vector of "shape" parameters of the distribution. In the case of the Multinom distribution, the model (as indicated above) has a rather different structure.

Except in the Gaussian case we assume that $\boldsymbol{\theta}$ is completely determined by a vector $\boldsymbol{x}$ of predictor variables and a vector $\boldsymbol{\phi}$ of predictor coefficients. We need to determine the first and second derivatives, of the likelihood of a single observation, with respect to the entries of $\boldsymbol{\phi}$, and in the case of the Gaussian distribution, with respect to the σ_i. We now

provide the details of the calculation of these derivatives for each of the five distributions in question.

2 The Gaussian Distribution

We denote the vector of standard deviations by $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_K)^\top$ (where K is the number of states). In the current development we assume that σ_i depends only on the state i of the underlying hidden Markov chain (and not on any other prectors included in $\boldsymbol{x}$. It is thus convenient to make explicit the dependence of the probability density functions upon the underlying state. We write the probability density function corresponding to state i as

$$f_i(y) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(\frac{-(y-\mu)^2}{2\sigma_i^2}\right) .$$

We model μ as $\mu = \boldsymbol{x}^\top \boldsymbol{\phi}$. Note that consequently μ depends, in general, upon the state i although this dependence $\boldsymbol{x}$ is not made explicit in the foregoing expression for $f_i(y)$. We need to differentiate $f_i(y)$ with respect to $\boldsymbol{\phi}$ and $\boldsymbol{\sigma}$.

It is straightforward, using logarithmic differentiation, to determine that:

$$\begin{aligned}
\frac{\partial f_i(y)}{\partial \mu} &= f_i(y)\left(\frac{y-\mu}{\sigma_i^2}\right) \\
\frac{\partial f_i(y)}{\partial \sigma_j} &= \begin{cases} f_i(y)\left(\frac{(y-\mu)^2}{\sigma_i^2} - 1\right)/\sigma_i & \text{if } j = i \\ 0 & \text{if } j \neq i \end{cases} \\
\frac{\partial^2 f_i(y)}{\partial \mu^2} &= f_i(y)\left(\frac{(y-\mu)^2}{\sigma_i^2} - 1\right)/\sigma_i^2 \\
\frac{\partial^2 f_i(y)}{\partial \sigma_i \partial \sigma_j} &= \begin{cases} f_i(y)\left(\left(\frac{(y-\mu)^2}{\sigma_i^2} - 1\right)^2 + 1 - \frac{3(y-\mu)^2}{\sigma_i^2}\right)/\sigma_i^2 & \text{if } j = i \\ 0 & \text{if } j \neq i \end{cases} \\
\frac{\partial^2 f_i(y)}{\partial \mu \partial \sigma_j} &= \begin{cases} f_i(y)\left(\frac{(y-\mu)^2}{\sigma^3} - \frac{3}{\sigma}\right)(y-\mu)/\sigma^2 & \text{if } j = i \\ 0 & \text{if } j \neq i \end{cases} .
\end{aligned} \tag{1}$$

Recalling that $\mu = \boldsymbol{x}^\top \boldsymbol{\phi}$ we see that

$$\frac{\partial \mu}{\partial \boldsymbol{\phi}} = \boldsymbol{x} ,$$

An application of the chain rule then gives:

$$\frac{\partial f_i(y)}{\partial \boldsymbol{\phi}} = \frac{\partial f_i(y)}{\partial \mu}\boldsymbol{x}$$

The second derivatives of $f_i(y)$ with respect to $\boldsymbol{\phi}$ are given by

$$\begin{aligned}\frac{\partial^2 f_i(y)}{\partial \boldsymbol{\phi}^\top \partial \boldsymbol{\phi}} &= \frac{\partial}{\partial \boldsymbol{\phi}^\top}\left(\frac{\partial f_i(y)}{\partial \mu}\boldsymbol{x}\right) \\ &= \boldsymbol{x}\left(\frac{\partial^2 f_i(y)}{\partial \mu^2}\frac{\partial \mu}{\partial \boldsymbol{\phi}^\top} + \frac{\partial^2 f_i(y)}{\partial \mu \partial \sigma_i}\frac{\partial \sigma_i}{\partial \boldsymbol{\phi}^\top}\right) \\ &= \left(\frac{\partial^2 f_i(y)}{\partial \mu^2}\right)\boldsymbol{x}\boldsymbol{x}^\top\end{aligned}$$

since $\partial \sigma_i / \partial \boldsymbol{\phi}^\top = \mathbf{0}$.

The second derivatives of $f_i(y)$ with respect to $\boldsymbol{\phi}$ and $\boldsymbol{\sigma}$ are given by

$$\frac{\partial^2 f_i(y)}{\partial \boldsymbol{\phi}^\top \partial \sigma_j} = \begin{cases} \left(\frac{\partial^2 f_i(y)}{\partial \mu \partial \sigma_j}\right)\boldsymbol{x}^\top & \text{if } j = i \\ \mathbf{0}^\top & \text{if } j \neq i \end{cases}$$

$$\frac{\partial^2 f_i(y)}{\partial \sigma_j \partial \boldsymbol{\phi}} = \begin{cases} \left(\frac{\partial^2 f_i(y)}{\partial \mu \partial \sigma_j}\right)\boldsymbol{x} & \text{if } j = i \\ \mathbf{0} & \text{if } j \neq i \end{cases}.$$

Note that

$$\frac{\partial^2 f_i(y)}{\partial \sigma_i \partial \sigma_j}$$

is provided in (1).

The structure of the first and second derivatives of $f_i(y)$ with respect to $\boldsymbol{\phi}$ and σ can be expressed concisely by letting

$$\boldsymbol{\psi} = \begin{bmatrix} \boldsymbol{\sigma} \\ \boldsymbol{\phi} \end{bmatrix}$$

and then writing

$$\begin{aligned}\frac{\partial f_i(y)}{\partial \boldsymbol{\psi}} &= \begin{bmatrix} \frac{\partial f_i(y)}{\partial \boldsymbol{\sigma}} \\ \frac{\partial f_i(y)}{\partial \boldsymbol{\phi}} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial f_i(y)}{\partial \sigma_i}\boldsymbol{\delta}_i \\ \frac{\partial f_i(y)}{\partial \mu}\boldsymbol{x} \end{bmatrix}\end{aligned}$$

where $\boldsymbol{\delta}_i$ is a vector of dimension K whose ith entry is 1 and whose other entries are all 0, and

$$\begin{aligned}\frac{\partial^2 f_i(y)}{\partial \boldsymbol{\psi}^\top \partial \boldsymbol{\psi}} &= \begin{bmatrix} \frac{\partial^2 f_i(y)}{\partial \boldsymbol{\sigma}^\top \partial \boldsymbol{\sigma}} & \frac{\partial^2 f_i(y)}{\partial \boldsymbol{\sigma}^\top \partial \boldsymbol{\phi}} \\ \frac{\partial^2 f_i(y)}{\partial \boldsymbol{\phi}^\top \partial \boldsymbol{\sigma}} & \frac{\partial^2 f_i(y)}{\partial \boldsymbol{\phi}^\top \partial \boldsymbol{\phi}} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial^2 f_i(y)}{\partial \sigma_i^2}\boldsymbol{\delta}_i\boldsymbol{\delta}_i^\top & \frac{\partial^2 f_i(y)}{\partial \mu \partial \sigma_i}\boldsymbol{\delta}_i\boldsymbol{x}^\top \\ \frac{\partial^2 f_i(y)}{\partial \mu \partial \sigma_i}\boldsymbol{x}\boldsymbol{\delta}_i^\top & \frac{\partial^2 f_i(y)}{\partial \mu^2}\boldsymbol{x}\boldsymbol{x}^\top \end{bmatrix}.\end{aligned}$$

Note that the firts and second partial derivatives of $f_i(y)$ with respect to μ and σ_i are provided in (1).

3 The Poisson Distribution

The likelihood is the probability mass function

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

$y = 0, 1, 2, \ldots$. Here $\boldsymbol{\theta}$ is a scalar, $\theta = \lambda$, and we model λ via $\lambda = \exp(\boldsymbol{x}^\top\boldsymbol{\phi})$, where $\boldsymbol{x}$ is a vector of predictors and $\boldsymbol{\phi}$ is a vector of predictor coefficients. The first and second derivatives of $f(y)$ with respect to λ are

$$\begin{aligned}\frac{\partial f(y)}{\partial \lambda} &= f(y)\left(\frac{y}{\lambda} - 1\right)\\ \frac{\partial^2 f(y)}{\partial \lambda^2} &= f(y)\left(\left(\frac{y}{\lambda} - 1\right)^2 - \frac{y}{\lambda^2}\right)\end{aligned}$$

Since $\lambda = \exp(\boldsymbol{x}^\top\boldsymbol{\phi})$ it follows readily that the first and second derivatives of λ with respect to $\boldsymbol{\phi}$ are $lambda\boldsymbol{x}$ and $\lambda\boldsymbol{x}\boldsymbol{x}^\top$, respectively. Applying the chain rule we get

$$\begin{aligned}\frac{\partial f(y)}{\partial \boldsymbol{\phi}} &= \frac{\partial f(y)}{\partial \lambda}\lambda\boldsymbol{x}\\ \frac{\partial^2 f(y)}{\partial \boldsymbol{\phi}^\top \partial \boldsymbol{\phi}} &= \left(\frac{\partial f(y)}{\partial \lambda}\lambda + \frac{\partial^2 f(y)}{\partial \lambda^2}\lambda^2\right)\boldsymbol{x}\boldsymbol{x}^\top\end{aligned}$$

4 The Binomial Distribution

The likelihood is the probability mass function

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

$y = 0, 1, 2, \ldots, n$, where n is the number of independent binomial trials on which the success count y is based, and p is the probability of success. Here $\boldsymbol{\theta}$ is a scalar, $\theta = p$, and we model p via $p = h(u)$ where $u = \boldsymbol{x}^\top\boldsymbol{\phi}$, where $\boldsymbol{x}$ is a vector of predictors, $\boldsymbol{\phi}$ is a vector of predictor coefficients and $h(u)$ is the logit function $h(u) = (1 + e^{-u})^{-1}$.

In what follows we will need the first and second derivatives of the logit function. These are given by

$$\begin{aligned}h'(u) &= \frac{e^{-u}}{(1+e^{-u})^2} \text{ and}\\ h''(u) &= \frac{e^{-u}(e^{-u}-1)}{(1+e^{-u})^3}\ .\end{aligned} \tag{2}$$

The first and second derivatives of $f(y)$ with respect to p are

$$\frac{\partial f(y)}{\partial p} = f(y)\left(\frac{y}{p} - \frac{n-y}{1-p}\right)$$
$$\frac{\partial^2 f(y)}{\partial p^2} = f(y)\left(\left(\frac{y}{p} - \frac{n-y}{1-p}\right)^2 - \frac{y}{p^2} - \frac{n-y}{(1-p)^2}\right) .$$

Since $p = h(\boldsymbol{x}^\top \boldsymbol{\phi})$ we see that

$$\frac{\partial p}{\partial \boldsymbol{\phi}} = h'(\boldsymbol{x}^\top \boldsymbol{\phi})\boldsymbol{x} \text{ and}$$
$$\frac{\partial^2 p}{\partial \boldsymbol{\phi}^\top \partial \boldsymbol{\phi}} = h''(\boldsymbol{x}^\top \boldsymbol{\phi})\boldsymbol{x}\boldsymbol{x}^\top$$

Applying the chain rule we see that

$$\frac{\partial f(y)}{\partial \boldsymbol{\phi}} = \frac{\partial f}{\partial p} h'(\boldsymbol{x}^\top \boldsymbol{\phi})\boldsymbol{x} \text{ and}$$
$$\frac{\partial^2 f(y)}{\partial \boldsymbol{\phi}^\top \partial \boldsymbol{\phi}} = \left(\frac{\partial f(y)}{\partial p} h''(\boldsymbol{x}^\top \boldsymbol{\phi}) + \frac{\partial^2 f(y)}{\partial p^2}(h'(\boldsymbol{x}^\top \boldsymbol{\phi})^2\right)\boldsymbol{x}\boldsymbol{x}^\top$$

Recall that expressions for $h'(\cdot)$ and $h''(\cdot)$ are given by (2).

5 The Db Distribution

The likelihood is the probability mass function which depends on a vector of parameters $\boldsymbol{\theta} = (\alpha, \beta)^\top$ and is somewhat complicated to write down. In order to obtain an expression for this probabilty mass function we need to define

$$\begin{aligned} h_0(y) &= (y(1-y))^{-1} \\ h(y) &= h_0((y - n_{\text{bot}} + 1)/(n_{\text{top}} - n_{\text{bot}} + 2)) \\ T_1(y) &= \log((y - n_{\text{bot}} + 1)/(n_{\text{top}} - n_{\text{bot}} + 2)) \\ T_2(y) &= \log((n_{\text{top}} - y + 1)/(n_{\text{top}} - n_{\text{bot}} + 2)) \\ A(\alpha, \beta) &= \log\left(\sum_{i=n_{\text{bot}}}^{n_{\text{top}}} h(i)\exp\{\alpha T_1(i) + \beta T_2(i)\}\right) . \end{aligned}$$

Given these definition the probability mass function of the Db distribution can be written as

$$f(y, \alpha, \beta) = \Pr(X = y \mid \alpha, \beta) = h(y)\exp\{\alpha T_1(y) + \beta T_2(y) - A(\alpha, \beta)\} .$$

We model α and β via

$$\alpha = \boldsymbol{x}^\top \boldsymbol{\phi}_1$$
$$\beta = \boldsymbol{x}^\top \boldsymbol{\phi}_2$$

where $\boldsymbol{x}$ is a vector of predictors and $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ are vectors of predictor coefficients. The vector $\boldsymbol{\phi}$, with respect to which we seek to differentiate the likelihood, is the catenation of $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$.

The first derivative of the likelihood with respect to $\boldsymbol{\phi}$ is

$$\begin{aligned}
\frac{\partial f}{\partial \boldsymbol{\phi}} &= \frac{\partial f}{\partial \alpha}\frac{\partial \alpha}{\partial \boldsymbol{\phi}} + \frac{\partial f}{\partial \beta}\frac{\partial \beta}{\partial \boldsymbol{\phi}} \\
&= \frac{\partial f}{\partial \alpha}\begin{bmatrix} \frac{\partial \alpha}{\partial \boldsymbol{\phi}_1} \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \frac{\partial \beta}{\partial \boldsymbol{\phi}_2} \end{bmatrix} \\
&= \frac{\partial f}{\partial \alpha}\begin{bmatrix} \boldsymbol{x} \\ \mathbf{0} \end{bmatrix} + \frac{\partial f}{\partial \beta}\begin{bmatrix} \mathbf{0} \\ \boldsymbol{x} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial f}{\partial \alpha}\boldsymbol{x} \\ \frac{\partial f}{\partial \beta}\boldsymbol{x} \end{bmatrix}
\end{aligned}$$

The second derivative is calculated as

$$\frac{\partial^2 f}{\partial \boldsymbol{\phi}^\top \partial \boldsymbol{\phi}} = \begin{bmatrix} \frac{\partial}{\partial \boldsymbol{\phi}^\top}\left(\frac{\partial f}{\partial \alpha}\boldsymbol{x}\right) \\ \frac{\partial}{\partial \boldsymbol{\phi}^\top}\left(\frac{\partial f}{\partial \beta}\boldsymbol{x}\right) \end{bmatrix}.$$

Taking this expression one row at a time we see that

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\phi}^\top}\left(\frac{\partial f}{\partial \alpha}\right) &= \begin{bmatrix} \frac{\partial}{\partial \boldsymbol{\phi}_1^\top}\left(\frac{\partial f}{\partial \alpha}\right) & \frac{\partial}{\partial \boldsymbol{\phi}_2^\top}\left(\frac{\partial f}{\partial \alpha}\right) \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial^2 f}{\partial \alpha^2}\frac{\partial \alpha}{\partial \boldsymbol{\phi}_1^\top} & \frac{\partial^2 f}{\partial \beta \partial \alpha}\frac{\partial \beta}{\partial \boldsymbol{\phi}_2^\top} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial^2 f}{\partial \alpha^2}\boldsymbol{x}^\top & \frac{\partial^2 f}{\partial \beta \partial \alpha}\boldsymbol{x}^\top \end{bmatrix} \text{ and likewise} \\
\frac{\partial}{\partial \boldsymbol{\phi}^\top}\left(\frac{\partial f}{\partial \beta}\right) &= \begin{bmatrix} \frac{\partial^2 f}{\partial \beta \partial \alpha}\boldsymbol{x}^\top & \frac{\partial^2 f}{\partial \beta^2}\boldsymbol{x}^\top \end{bmatrix}.
\end{aligned}$$

Combining the foregoing we get

$$\frac{\partial^2 f}{\partial \boldsymbol{\phi}^\top \partial \boldsymbol{\phi}} = \begin{bmatrix} \frac{\partial^2 f}{\partial \alpha^2}\boldsymbol{x}\boldsymbol{x}^\top & \frac{\partial^2 f}{\partial \beta \partial \alpha}\boldsymbol{x}\boldsymbol{x}^\top \\ \frac{\partial^2 f}{\partial \beta \partial \alpha}\boldsymbol{x}\boldsymbol{x}^\top & \frac{\partial^2 f}{\partial \beta^2}\boldsymbol{x}\boldsymbol{x}^\top \end{bmatrix}.$$

As was the case for the three distributions for which $\boldsymbol{\theta}$ is a scalar, it is expedient to express the partial derivatives of $f(y, \alpha, \beta)$, with respect to the parameters of the distribution, in

terms of $f(y, \alpha, \beta)$ The required expressions are as follows:

$$\begin{aligned}
\frac{\partial f}{\partial \alpha} &= f(y, \alpha, \beta) \left(T_1(y) - \frac{\partial A}{\partial \alpha} \right) \\
\frac{\partial f}{\partial \beta} &= f(y, \alpha, \beta) \left(T_2(y) - \frac{\partial A}{\partial \beta} \right) \\
\frac{\partial^2 f}{\partial \alpha^2} &= f(y, \alpha, \beta) \left[\left(T_1(y) - \frac{\partial A}{\partial \alpha} \right)^2 - \frac{\partial^2 A}{\partial \alpha^2} \right] \\
\frac{\partial^2 f}{\partial \alpha \partial \beta} &= f(y, \alpha, \beta) \left[\left(T_1(y) - \frac{\partial A}{\partial \alpha} \right) \left(T_2(y) - \frac{\partial A}{\partial \beta} \right) - \frac{\partial^2 A}{\partial \alpha \partial \beta} \right] \\
\frac{\partial^2 f}{\partial \beta^2} &= f(y, \alpha, \beta) \left[\left(T_2(y) - \frac{\partial A}{\partial \beta} \right)^2 - \frac{\partial^2 A}{\partial \beta^2} \right]
\end{aligned}$$

It remains to provide expressions for the partial derivatives of A with respect to α and β. Let

$$E = \exp(A) = \sum_{i=n_{\text{bot}}}^{n_{\text{top}}} h(i) \exp\{\alpha T_1(i) + \beta T_2(i)\} .$$

Clearly

$$\begin{aligned}
\frac{\partial A}{\partial \alpha} &= \frac{1}{E}\frac{\partial E}{\partial \alpha} \\
\frac{\partial A}{\partial \beta} &= \frac{1}{E}\frac{\partial E}{\partial \beta} \\
\frac{\partial^2 A}{\partial \alpha^2} &= \frac{1}{E}\frac{\partial^2 E}{\partial \alpha^2} - \frac{1}{E^2}\left(\frac{\partial E}{\partial \alpha}\right)^2 \\
\frac{\partial^2 A}{\partial \alpha \partial \beta} &= \frac{1}{E}\frac{\partial^2 E}{\partial \alpha \partial \beta} - \frac{1}{E^2}\left(\frac{\partial E}{\partial \alpha}\frac{\partial E}{\partial \beta}\right) \\
\frac{\partial^2 A}{\partial \beta^2} &= \frac{1}{E}\frac{\partial^2 E}{\partial \beta^2} - \frac{1}{E^2}\left(\frac{\partial E}{\partial \beta}\right)^2
\end{aligned}$$

Finally, the relevant partial derivatives of E are:

$$\begin{aligned}
\frac{\partial E}{\partial \alpha} &= \sum_{i=n_{\text{bot}}}^{n_{\text{top}}} h(i) T_1(i) \exp(\alpha T_1(i) + \beta T_2(i)) \\
\frac{\partial E}{\partial \beta} &= \sum_{i=n_{\text{bot}}}^{n_{\text{top}}} h(i) T_2(i) \exp(\alpha T_1(i) + \beta T_2(i)) \\
\frac{\partial^2 E}{\partial \alpha^2} &= \sum_{i=n_{\text{bot}}}^{n_{\text{top}}} h(i) T_1(i)^2 \exp(\alpha T_1(i) + \beta T_2(i)) \\
\frac{\partial^2 E}{\partial \alpha \partial \beta} &= \sum_{i=n_{\text{bot}}}^{n_{\text{top}}} h(i) T_1(i) T_2(i) \exp(\alpha T_1(i) + \beta T_2(i)) \\
\frac{\partial^2 E}{\partial \beta^2} &= \sum_{i=n_{\text{bot}}}^{n_{\text{top}}} h(i) T_2(i)^2 \exp(\alpha T_1(i) + \beta T_2(i)) .
\end{aligned}$$

6 The Multinom Distribution

This distribution is very different from those with which we have previously dealt. It is defined effectively in terms of *tables*. In the hidden Markov model context, these tables take the form

$$\Pr(Y = y_i \mid S = k) = \rho_{ik}$$

where Y is the emissions variate, its possible values or "levels" are $y_1, y_2, \ldots, y_m$, and S denotes "state" which (wlog) takes values $1, 2, \ldots, K$. Of course $\rho_{\cdot k} = 1$ for all k. We shall

denote $\Pr(Y = y \mid S = k) = \rho_{ik}$ by $f_k(y)$. Thus instead of having a single probability mass function, we have K of them.

The maximisation of the likelihood with respect to the ρ_{ik} is awkward, due to the forgoing "sum-to-1" constraint, and it is better to impose this constraint "smoothly" via a logistic parameterisation. Such a parameterisation also allows us to express the dependence upon "state" in terms of linear predictors, which opens up the possibility of including predictors, other than those determined by "state", in the model.

To this end we define vectors of parameters $\boldsymbol{\phi}_i$, $i = 1, \ldots, m$, corresponding to each of the possible values of Y. For identifiability we take $\boldsymbol{\phi}_m$ to be identically 0. Each $\boldsymbol{\phi}_i$ is a vector of length np, say, where np is the number of predictors. If, in a K state model, there are no predictors other than those determined by state, then $np = K$. In this case there are $K \times (m-1)$ "free" parameters, just as there should be (and just at there are in the original parameterisation in terms of the ρ_{ik}). Let the kth entry of $\boldsymbol{\phi}_i$ be ϕ_{ik}, $k = 1, \ldots, np$. Let $\boldsymbol{\phi}$ be the vector consisting of the catenation of all of the ϕ_{ij}, excluding the entries of $\boldsymbol{\phi}_m$ which are all 0:

$$\boldsymbol{\phi} = (\phi_{11}, \phi_{12}, \ldots, \phi_{1,np}, \phi_{21}, \phi_{22}, \ldots, \phi_{2,np}, \cdots \ , \cdots \ , \phi_{m-1,1}, \phi_{m-1,2}, \ldots, \phi_{m-1,np})^\top \ .$$

Let $\boldsymbol{x}$ be a vector of predictors. In terms of the foregoing notation, $f_k(y)$ can be written as

$$f_k(y) = \frac{e^{\boldsymbol{x}^\top \boldsymbol{\phi}_y}}{Z}$$

where in turn

$$Z = \sum_{\ell=1}^{k} e^{\boldsymbol{x}^\top \boldsymbol{\phi}_\ell} \ .$$

The dependence of $f_k(y)$ upon the state k is incorporated in the predictor vector $\boldsymbol{x}$ which includes predictors indicating state. We now calculate the partial derivatives of $f_k(y)$ with respect to $\boldsymbol{\phi}$. First note that $\frac{\partial f}{\partial \boldsymbol{\phi}}$ can be written as

$$\begin{bmatrix} \frac{\partial f_k}{\partial \boldsymbol{\phi}_1} \\ \frac{\partial f_k}{\partial \boldsymbol{\phi}_2} \\ \vdots \\ \frac{\partial f_k}{\partial \boldsymbol{\phi}_{m-1}} \end{bmatrix} .$$

Next we calculate

$$\frac{\partial f_k(y)}{\partial \boldsymbol{\phi}_i}, \quad i = 1, \ldots, m-1 \ .$$

Using logarithmic differentiation we see that

$$\frac{1}{f_k(y)} \frac{\partial f_k(y)}{\partial \boldsymbol{\phi}_i} = \delta_{yi} \boldsymbol{x} - \frac{1}{Z} e^{\boldsymbol{x}^\top \boldsymbol{\phi}_i} \boldsymbol{x}$$

so that

$$\frac{\partial f_k(y)}{\partial \boldsymbol{\phi}_i} = f_k(y)\left(\delta_{yi} - \frac{e^{\boldsymbol{x}^\top \boldsymbol{\phi}_i}}{Z}\right)$$

which can be written as $f_k(y)(\delta_{yi} - f_k(i))\boldsymbol{x}$.

In summary we have

$$\frac{\partial f}{\partial \boldsymbol{\phi}} = f_k(y)\begin{bmatrix} (\delta_{y1} - f_k(1))\boldsymbol{x} \\ (\delta_{y2} - f_k(2))\boldsymbol{x} \\ \vdots \\ (\delta_{y,m-1} - f_k(m-1))\boldsymbol{x} \end{bmatrix}$$

The second derivatives of $f_k(y)$ with respect to $\boldsymbol{\phi}$ are given by

$$\frac{\partial^2 f}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}^\top} = \begin{bmatrix} \frac{\partial^2 f}{\partial \boldsymbol{\phi}_1 \partial \boldsymbol{\phi}_1^\top} & \frac{\partial^2 f}{\partial \boldsymbol{\phi}_1 \partial \boldsymbol{\phi}_2^\top} & \cdots & \frac{\partial^2 f}{\partial \boldsymbol{\phi}_1 \partial \boldsymbol{\phi}_{m-1}^\top} \\ \frac{\partial^2 f}{\partial \boldsymbol{\phi}_2 \partial \boldsymbol{\phi}_1^\top} & \frac{\partial^2 f}{\partial \boldsymbol{\phi}_2 \partial \boldsymbol{\phi}_2^\top} & \cdots & \frac{\partial^2 f}{\partial \boldsymbol{\phi}_2 \partial \boldsymbol{\phi}_{m-1}^\top} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 f}{\partial \boldsymbol{\phi}_{m-1} \partial \boldsymbol{\phi}_1^\top} & \frac{\partial^2 f}{\partial \boldsymbol{\phi}_{m-1} \partial \boldsymbol{\phi}_2^\top} & \cdots & \frac{\partial^2 f}{\partial \boldsymbol{\phi}_{m-1} \partial \boldsymbol{\phi}_{m-1}^\top} \end{bmatrix}$$

The (i,j)th entry of $\frac{\partial^2 f}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}^\top}$, i.e. $\frac{\partial^2 f}{\partial \boldsymbol{\phi}_i \partial \boldsymbol{\phi}_j^\top}$, is given by

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\phi}_i}\left(\frac{\partial y}{\partial \boldsymbol{\phi}_j^\top}\right) &= \frac{\partial}{\partial \boldsymbol{\phi}_i}\left(f_k(y)(\delta_{yj} - f_k(j)\boldsymbol{x}^\top\right) \\ &= f_k(y)(0 - f_k(j)(\delta_{ij} - f_k(i))\boldsymbol{x}\boldsymbol{x}^\top) + f_k(y)(\delta_{yi} - f_k(i))\boldsymbol{x}(\delta_{yj} - f_k(j))\boldsymbol{x}^\top \\ &= f_k(y)(-f_k(j)(\delta_{ij} - f_k(i)) + (\delta_{yj} - f_k(i))(\delta_{yj} - f_k(j)))\boldsymbol{x}\boldsymbol{x}^\top \\ &= f_k(y)(f_k(i)(f_k(j) - \delta_{ij} f_k(j) + (\delta_{yi} - f_k(i))(\delta_{yj} - f_k(j)))\boldsymbol{x}\boldsymbol{x}^\top \end{aligned}$$

At first glance this expression seems to be anomalously asymmetric in i and j, but the asymmetry is illusory. Note that when $i \neq j$, $\delta_{ij} f_k(j)$ is 0, and when $i = j$, $\delta_{ij} f_k(j) = f_k(j) = f_k(i)$.

In summary we see that

$$\frac{\partial^2 f}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}^\top} = \begin{bmatrix} a_{11}\boldsymbol{x}\boldsymbol{x}^\top & a_{12}\boldsymbol{x}\boldsymbol{x}^\top & \dots & a_{1,m-1}\boldsymbol{x}\boldsymbol{x}^\top \\ a_{21}\boldsymbol{x}\boldsymbol{x}^\top & a_{22}\boldsymbol{x}\boldsymbol{x}^\top & \dots & a_{2,m-1}\boldsymbol{x}\boldsymbol{x}^\top \\ \vdots & \vdots & \vdots & \vdots \\ a_{m-1,1}\boldsymbol{x}\boldsymbol{x}^\top & a_{m-1,2}\boldsymbol{x}\boldsymbol{x}^\top & \dots & a_{m-1,,m-1}\boldsymbol{x}\boldsymbol{x}^\top \end{bmatrix}$$

where $a_{ij} = f_k(y)(f_k(i)(f_k(j) - \delta_{ij} f_k(j) + (\delta_{yi} - f_k(i))(\delta_{yj} - f_k(j))$, $i, j = 1, \dots, m-1$.