One observes a sample in which p-dimensional random vectors $\{\mathbf{x}_{jk}\}_{k=1}^{n_j}$ arise from population $j \in \{1, ..., J\}$. We wish to fit a multivariate logistic model, for which the conditional probability of membership in class j is given by

$$p(j|\mathbf{x}) = \frac{e^{\alpha_j + \boldsymbol{\beta}_j^T \mathbf{x}}}{1 + \sum_{j=1}^{J-1} e^{\alpha_j + \boldsymbol{\beta}_j^T \mathbf{x}}}, \ j = 1, ..., J-1,$$
$$p(J|\mathbf{x}) = \frac{1}{1 + \sum_{j=1}^{J-1} e^{\alpha_j + \boldsymbol{\beta}_j^T \mathbf{x}}} = 1 - p(1|\mathbf{x}) - \cdots - p(J-1|\mathbf{x}).$$

Define $d = (p+1)(J-1)$ and

$$\mathbf{z}_{jk}|_{(p+1)\times 1} = \begin{pmatrix} 1 \\ \mathbf{x}_{jk} \end{pmatrix}, \ \boldsymbol{\theta}_j|_{(p+1)\times 1} = \begin{pmatrix} \alpha_j \\ \boldsymbol{\beta}_j \end{pmatrix},$$
$$\mathbf{Z}_{jk}|_{(p+1)\times(p+1)} = \mathbf{z}_{jk}\mathbf{z}_{jk}^T, \ \boldsymbol{\theta}|_{d\times 1} = \begin{pmatrix} \boldsymbol{\theta}_1 \\ \vdots \\ \boldsymbol{\theta}_{J-1} \end{pmatrix}.$$

Then $p(j|\mathbf{x}) = e^{\boldsymbol{\theta}_j^T \mathbf{z}} p(J|\mathbf{x})$ for $j < J$ and the log-likelihood is

$$l(\boldsymbol{\theta}) = \sum_{j=1}^{J}\sum_{k=1}^{n_j} \log p(j|\mathbf{x}_{jk}) = \sum_{j=1}^{J-1}\sum_{k=1}^{n_j} \boldsymbol{\theta}_j^T \mathbf{z}_{jk} + \sum_{j=1}^{J}\sum_{k=1}^{n_j} \log p(J|\mathbf{x}_{jk}).$$

1. The gradient of l is the $d \times 1$ vector

   $$\dot{l}(\boldsymbol{\theta}) = \mathbf{t} - \sum_{j=1}^{J}\sum_{k=1}^{n_j} \left(\mathbf{p}(\mathbf{x}_{jk}|\boldsymbol{\theta}) \otimes \mathbf{z}_{jk}\right),$$

   where

   $$\mathbf{p}(\mathbf{x}|\boldsymbol{\theta}) = \begin{pmatrix} p(1|\mathbf{x}) \\ \vdots \\ p(J-1|\mathbf{x}) \end{pmatrix} : (J-1) \times 1$$

   and where

   $$\mathbf{t} = \begin{pmatrix} \sum_{k=1}^{n_j} \mathbf{z}_{1k} \\ \vdots \\ \sum_{k=1}^{n_j} \mathbf{z}_{J-1,k} \end{pmatrix} : d \times 1$$

   is the vector of totals.

2. The Hessian is the $d \times d$ matrix

   $$\ddot{l}(\boldsymbol{\theta}) = -\sum_{j=1}^{J}\sum_{k=1}^{n_j} \left(\mathbf{W}(\mathbf{x}_{jk}|\boldsymbol{\theta}) \otimes \mathbf{Z}_{jk}\right),$$

   where

   $$\mathbf{W}(\mathbf{x}_{jk}|\boldsymbol{\theta}) = \begin{pmatrix} p(1|\mathbf{x}_{jk}) & & 0 \\ & \ddots & \\ 0 & & p(J-1|\mathbf{x}_{jk}) \end{pmatrix} - \mathbf{p}(\mathbf{x}_{jk}|\boldsymbol{\theta})\,\mathbf{p}^T(\mathbf{x}_{jk}|\boldsymbol{\theta}).$$

Thus the Newton-Raphson iterates are

$$\boldsymbol{\theta}_{m+1} = \boldsymbol{\theta}_m + \left(\sum_{j=1}^{J}\sum_{k=1}^{n_j} \left(\mathbf{W}(\mathbf{x}_{jk}|\boldsymbol{\theta}_m) \otimes \mathbf{Z}_{jk}\right)\right)^{-1} \left(\mathbf{t} - \sum_{j=1}^{J}\sum_{k=1}^{n_j} \left(\mathbf{p}(\mathbf{x}_{jk}|\boldsymbol{\theta}_m) \otimes \mathbf{z}_{jk}\right)\right). \quad (1)$$

Note that the Hessian can be estimated from the sample proportions. If $\mathbf{w} = (n_1/n, \cdots, n_{J-1}/n)^T$ then $\mathbf{W}(\mathbf{x}_{jk}|\boldsymbol{\theta}_m)$ can be estimated by the constant matrix

$$\hat{\mathbf{W}} = diag(\mathbf{w}) - \mathbf{w}\mathbf{w}^T,$$

and then (1) becomes

$$\boldsymbol{\theta}_{m+1} = \boldsymbol{\theta}_m + \left(\hat{\mathbf{W}}^{-1} \otimes \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\right)\left(\mathbf{t} - \sum_{j=1}^{J}\sum_{k=1}^{n_j}\left(\mathbf{p}(\mathbf{x}_{jk}|\boldsymbol{\theta}_m) \otimes \mathbf{z}_{jk}\right)\right).$$

(This turns out to be *very* slow; on the other hand the Hessian is often nearly singular.)