Information Principles in Random Effects Models

Herwig Friedl Göran Kauermann

August 14, 2013

1 Information Matrices

In this note, data (y, z) are considered, where y denotes the observable part and z refers to that part which is unobservable. Later we will concentrate on a response variable y which is modelled in terms of a fixed predictor variable and an additional random effect z.

1.1 Complete Likelihood

Let θ denote all unknown parameters in the model and

$$\lambda_c(y, z; \theta) = \log f(y, z; \theta)$$

be the complete log-likelihood function corresponding to the joint distribution of the response and the random effect. We consider the model

$$f(y, z; \theta) = f(y|z; \theta) \times f(z; \theta)$$

where $f(y|z; \theta)$ is the conditional model, given the random effect z, and $f(z; \theta)$ denotes the random effect density. Both, $f(y|z; \theta)$ and $f(z; \theta)$ possibly depend on unknown parameters θ.

The respective complete score vector is

$$\begin{aligned} S_c(y, z; \theta) &= \frac{\partial}{\partial \theta} \lambda_c(y, z; \theta) \\ &= \frac{f'(y, z; \theta)}{f(y, z; \theta)} \qquad (1) \end{aligned}$$

with complete negative second derivative

$$\begin{aligned} I_c(y, z; \theta) &= -\frac{\partial^2}{\partial \theta \partial \theta^t} \lambda_c(y, z; \theta) \\ &= -\frac{f''(y, z; \theta)}{f(y, z; \theta)} + \frac{f'(y, z; \theta)}{f(y, z; \theta)} \frac{f'^t(y, z; \theta)}{f(y, z; \theta)} \\ &= -\frac{f''(y, z; \theta)}{f(y, z; \theta)} + S_c(y, z; \theta) S_c^t(y, z; \theta). \qquad (2) \end{aligned}$$

1.2 Observed Likelihood

The MLE $\hat{\theta}$ is constructed by maximizing the observed log-likelihood

$$\lambda(y;\theta) = \log \int f(y,z;\theta)dz$$

which depends on the observations y only. In what follows we need the exchangeability of integration and differentiation of the complete density function. From the respective observed scores

$$\begin{aligned} S(y;\theta) &= \frac{\partial}{\partial\theta}\lambda(y;\theta) \\ &= \frac{\int f'(y,z;\theta)dz}{\int f(y,z;\theta)dz} \\ &= \int \frac{f'(y,z;\theta)}{f(y,z;\theta)} \frac{f(y,z;\theta)}{\int f(y,z;\theta)dz}dz \\ &= E_\theta\left(S_c(y,z;\theta)|y\right) \qquad (3) \end{aligned}$$

we get the matrix of observed negative second derivatives

$$\begin{aligned} I(y;\theta) &= -\frac{\partial^2}{\partial\theta\partial\theta^t}\lambda(y;\theta) \\ &= -\frac{\int f''(y,z;\theta)dz}{\int f(y,z;\theta)dz} + \frac{\int f'(y,z;\theta)dz}{\int f(y,z;\theta)dz}\frac{\int f'^t(y,z;\theta)dz}{\int f(y,z;\theta)dz} \\ &= -\int \frac{f''(y,z;\theta)}{f(y,z;\theta)}\frac{f(y,z;\theta)}{\int f(y,z;\theta)dz}dz + S(y;\theta)S^t(y;\theta) \\ &= E_\theta\left(-\frac{f''(y,z;\theta)}{f(y,z;\theta)}\Big|y\right) + S(y;\theta)S^t(y;\theta). \end{aligned}$$

Because $-f''(y,z;\theta)/f(y,z;\theta) = I_c(y,z;\theta) - S_c(y,z;\theta)S_c^t(y,z;\theta)$ from (2), this can be rewritten as

$$\begin{aligned} I(y;\theta) &= E_\theta\left(I_c(y,z;\theta)|y\right) - E_\theta\left(S_c(y,z;\theta)S_c^t(y,z;\theta)|y\right) + E_\theta(S_c(y,z;\theta)|y)E_\theta(S_c^t(y,z;\theta)|y) \\ &= E_\theta\left(I_c(y,z;\theta)|y\right) - Var_\theta\left(S_c(y,z;\theta)|y\right) \qquad (4) \end{aligned}$$

often called the missing information principle. This result is due to Louis (1982) and used to extract the observed information matrix when the EM algorithm is applied to find MLEs in incomplete data problems. Moreover, it provides a means of estimating the information which is associated with the MLEs and requires only the computation of a complete-data gradient vector and a second derivative matrix but not those associated with the incomplete-data likelihood.

1.3 Missing Information Principle

Denote the observed and complete Fisher information matrices by

$$\begin{aligned} \mathcal{I}(\theta) &= E_\theta(I(y;\theta)) = E_\theta(S(y;\theta)S^t(y;\theta)) \\ \mathcal{I}_c(\theta) &= E_\theta(I_c(y,z;\theta)) = E_\theta(S_c(y,z;\theta)S_c^t(y,z;\theta)), \end{aligned}$$

where expectations are taken over the marginal and the joint density, respectively. From

$$\lambda(y;\theta) = \lambda_c(y,z;\theta) - \log f(z|y;\theta)$$

we get the identity

$$\begin{aligned} -\frac{\partial^2}{\partial\theta\partial\theta^t}\lambda(y;\theta) &= -\frac{\partial^2}{\partial\theta\partial\theta^t}\lambda_c(y,z;\theta) + \frac{\partial^2}{\partial\theta\partial\theta^t}\log f(z|y;\theta) \\ I(y;\theta) &= I_c(y,z;\theta) + \frac{\partial^2}{\partial\theta\partial\theta^t}\log f(z|y;\theta). \end{aligned}$$

By taking the expectation with respect to the conditional density of z, given the data y, we have

$$I(y;\theta) = E_\theta(I_c(y,z;\theta)|y) - E_\theta\left(-\frac{\partial^2}{\partial\theta\partial\theta^t}\log f(z|y;\theta)\Big| y\right).$$

McLachlan and Krishnan (1997) define

$$\begin{aligned} \mathcal{I}_c(y;\theta) &:= E_\theta\left(I_c(y,z;\theta)|y\right) \\ \mathcal{I}_m(y;\theta) &:= E_\theta\left(-\frac{\partial^2}{\partial\theta\partial\theta^t}\log f(z|y;\theta)\Big| y\right) \end{aligned}$$

to rewrite the above identity as a difference of conditionally expected informations, namely

$$I(y;\theta) = \mathcal{I}_c(y;\theta) - \mathcal{I}_m(y;\theta), \tag{5}$$

where $\mathcal{I}_m(y;\theta)$ is considered as the missing information as a consequence of observing y only and not also z. As stated in Louis (1982) the comparison of (5) with (4) results in an easy to interpret expression for the missing information, i.e.

$$\mathcal{I}_m(y;\theta) = Var_\theta(S_c(y,z;\theta)|y). \tag{6}$$

Note that with respect to the marginal density we have

$$E_\theta(\mathcal{I}_c(y;\theta)) = \int\left(\int I_c(y,z;\theta)f(z|y;\theta)\ dz\right)f(y;\theta)\ dy = \int\int I_c(y,z;\theta)f(y,z;\theta)\ dy\ dz = \mathcal{I}_c(\theta).$$

By taking marginal expectations in (5) and by using the above result, we get

$$\mathcal{I}(\theta) = \mathcal{I}_c(\theta) - E_\theta(\mathcal{I}_m(y;\theta)) \tag{7}$$

for the a priori expected information.

1.4 EM Estimates

Dempster, Laird and Rubin (1977) iteratively maximizes

$$Q(\theta|\theta^{(t)}) = \int \lambda_c(y,z;\theta) f(z|y;\theta^{(t)})dz = E_{\theta^{(t)}}(\lambda_c(y,z;\theta)|y),$$

the conditional expected joint log-likelihood given the data. They also showed that if $\theta^{(t)}$ converges to a point $\hat{\theta}$ then the observed score function has a zero at $\theta^{(\infty)} = \hat{\theta}$, the marginal MLE. Each iteration of the EM algorithm solves

$$\frac{\partial}{\partial\theta} Q(\theta|\theta^{(t)})\bigg|_{\theta=\theta^{(t+1)}} = 0. \tag{8}$$

Louis (1982) and later on Meilijson (1989) considered Q and its derivatives when evaluated in $\theta = \theta^{(t)}$, e.g.

$$\begin{aligned}
Q(\theta|\theta^{(t)})\bigg|_{\theta=\theta^{(t)}} &= E_{\theta^{(t)}}(\lambda_c(y,z;\theta^{(t)})|y) \\
\frac{\partial}{\partial\theta} Q(\theta|\theta^{(t)})\bigg|_{\theta=\theta^{(t)}} &= E_{\theta^{(t)}}(S_c(y,z;\theta^{(t)})|y) = S(y;\theta^{(t)}) \\
-\frac{\partial^2}{\partial\theta\partial\theta^t} Q(\theta|\theta^{(t)})\bigg|_{\theta=\theta^{(t)}} &= E_{\theta^{(t)}}(I_c(y,z;\theta^{(t)})|y) = \mathcal{I}_c(y;\theta^{(t)}).
\end{aligned}$$

Evaluating the missing information (6) at $\hat{\theta}$ gives us the simplification

$$\mathcal{I}_m(y;\hat{\theta}) = E_\theta(S_c(y,z;\theta)S_c^t(y,z;\theta)|y)|_{\theta=\hat{\theta}} \tag{9}$$

because $E_\theta(S_c(y,z;\theta)|y)\big|_{\theta=\hat{\theta}} = S(y;\hat{\theta}) = 0$ at convergence, a consequence of (8).

A software which provides EM estimates by directly applying successive E and M steps will automatically provide the inverse of the matrix $\mathcal{I}_c(y;\hat{\theta})$, an estimate of the complete information at convergence. But this matrix should not be used for estimating the standard errors of the MLE $\hat{\theta}$, because it does not account for the missing information. Instead of that, Efron and Hinkley (1978) suggest to use the inverse of the observed information $I(y;\hat{\theta})$ to serve as an estimate of the covariance matrix of the MLE.

If we take the observed score vector $S(y;\theta)$ and calculate its negative derivative then we get the desired observed information

$$-\frac{\partial}{\partial\theta} S(y;\theta) = I(y;\theta).$$

The Newton-Raphson procedure

$$\theta^{(t+1)} = \theta^{(t)} + I(y;\theta^{(t)})^{-1} S(y;\theta^{(t)})$$

will give $I(y;\hat{\theta})^{-1}$ at convergence. One remaining open problem is concerned with the question on how we can compute or approximate all the expected values that we need.

2 A Model to Handle Overdispersion

Now we consider a Generalized Linear Model for the conditional mean $\mu = \mu(z) = E(y|z;\beta)$ of an observation y given an unobservable random effect z of the form

$$g(\mu) = x^t\beta + z$$

where x includes all explanatory variables, y is a response variable and z is the random effect. β is the vector of unknown parameters associated to the fixed effects and we like to construct estimates for the standard error of the maximum likelihood estimate (MLE) $\hat{\beta}$. We also assume that distributional assumptions can be made on the response variable conditional on the random effect.

2.1 Gaussian Random Effects

First let us assume that data $y = (y_1, \ldots, y_n)^t$ are available that conditionally follow a Generalized Linear Model where the random effects $z = (z_1, \ldots, z_n)^t$ are independently drawn from a normal distribution with zero mean and variance σ_z^2, e.g.

$$g(\mu_i) = x_i^t\beta + \sigma_z z_i,$$

where g is a (canonical) link function and $z_i \overset{iid}{\sim} N(0,1)$. Here, $\theta = (\beta^t, \sigma_z)^t$ and the density of the random effects $f(z) = \varphi(z)$ does not depend on any unknown parameter. Let $\tilde{X} = (X|z)$ denote the design matrix X extended by the vector of random effects z and write

$$g(\mu) = \tilde{X}\theta.$$

For $f(y|z;\theta)$ from the exponential family and a canonical link model we have the well known results

$$\begin{aligned}
S_c(y,z;\theta) &= \frac{f'(y|z;\theta) \times \varphi(z)}{f(y|z;\theta) \times \varphi(z)} = \frac{f'(y|z;\theta)}{f(y|z;\theta)} = \frac{\partial}{\partial\theta}\log f(y|z;\theta) \\
&= \tilde{X}^t(y-\mu) \\
I_c(y,z;\theta) &= -\frac{\partial^2}{\partial\theta\partial\theta^t}\log f(y|z;\theta) \\
&= \tilde{X}^t V \tilde{X}.
\end{aligned}$$

Here, $V = V(\mu(z))$ denotes the conditional variance function, i.e. the variance function to $f(y|z;\theta)$ as in the usual GLM setting for $y|z$. Therefore, the information matrices in (5) are

$$\begin{aligned}
\mathcal{I}_c(y;\theta) &= E_\theta\Big(I_c(y,z;\theta)\Big|y\Big) = E_\theta\Big(\tilde{X}^t V \tilde{X}\Big|y\Big) \\
\mathcal{I}_m(y;\theta) &= E_\theta\Big(S_c(y,z;\theta)S_c^t(y,z;\theta)\Big|y\Big) - E_\theta\Big(S_c(y,z;\theta)\Big|y\Big)E_\theta\Big(S_c^t(y,z;\theta)\Big|y\Big) \\
&= E_\theta\Big(\tilde{X}^t(y-\mu)(y-\mu)^t\tilde{X}\Big|y\Big) - E_\theta\Big(\tilde{X}^t(y-\mu)\Big|y\Big)E_\theta\Big((y-\mu)^t\tilde{X}\Big|y\Big).
\end{aligned}$$

We approximate the above conditional expectations through Gauss-Hermite quadrature. That means that the unobservable effects z_i are replaced by some known masspoints ζ_k with known masses π_k, $k = 1, \ldots, K$. Further, let $\tilde{x}_{ik} = (x_i^t, \zeta_k)^t$ and

$$w_{ik} = \frac{f(y_i|\zeta_k;\theta)\pi_k}{\sum_{l=1}^K f(y_i|\zeta_l;\theta)\pi_l}$$

the respective approximation to $f(z_i|y_i;\theta)$. This gives an approximation to the **complete Fisher information**

$$\begin{aligned}\mathcal{I}_c(y;\theta) &= \sum_i E_\theta(\tilde{x}_i\tilde{x}_i^t V_i|y_i) \\ &\approx \sum_i\sum_k \tilde{x}_{ik}\tilde{x}_{ik}^t V_{ik} w_{ik},\end{aligned}$$

which is automatically provided at convergence of the EM algorithm. For the first term in $\mathcal{I}_m(y;\theta)$ we get

$$\begin{aligned}&E_\theta\left(\sum_i \tilde{x}_i\tilde{x}_i^t(y_i-\mu_i)^2 + \sum_{i\neq j}\tilde{x}_i\tilde{x}_j^t(y_i-\mu_i)(y_j-\mu_j)\Big| y\right) \\ \approx &\sum_i\sum_k \tilde{x}_{ik}\tilde{x}_{ik}^t(y_i-\mu_{ik})^2 w_{ik} + \sum_{i\neq j}\sum_k\sum_l \tilde{x}_{ik}\tilde{x}_{jl}^t(y_i-\mu_{ik})(y_j-\mu_{jl})w_{ik}w_{jl}.\end{aligned}$$

The second term of $\mathcal{I}_m(y;\theta)$ is zero at $\theta = \hat{\theta}$. Generally, it can be approximated by

$$E_\theta\left(\sum_i \tilde{x}_i(y_i-\mu_i)\Big| y\right) E_\theta\left(\sum_j \tilde{x}_j^t(y_j-\mu_j)\Big| y\right) \approx \sum_i\sum_k \tilde{x}_{ik}(y_i-\mu_{ik})w_{ik}\sum_j\sum_l \tilde{x}_{jl}^t(y_j-\mu_{jl})w_{jl}.$$

Subtracting the last from the previous result gives the approximation to the **missing information**

$$\begin{aligned}\mathcal{I}_m(y;\theta) &= \sum_i E_\theta\left(\tilde{x}_i\tilde{x}_i^t(y_i-\mu_i)^2|y_i\right) - \sum_i E_\theta\left(\tilde{x}_i(y_i-\mu_i)|y_i\right)E_\theta\left(\tilde{x}_i^t(y_i-\mu_i)|y_i\right) \\ &\approx \sum_i\sum_k \tilde{x}_{ik}\tilde{x}_{ik}^t(y_i-\mu_{ik})^2 w_{ik} - \sum_i\sum_k\sum_l \tilde{x}_{ik}\tilde{x}_{il}^t(y_i-\mu_{ik})(y_i-\mu_{il})w_{ik}w_{il}.\end{aligned}$$

Hence, the **observed information** is

$$\begin{aligned}I(y;\theta) &= \sum_i E_\theta(\tilde{x}_i x_i^t V_i|y_i) \\ &\quad -\sum_i E_\theta\left(\tilde{x}_i\tilde{x}_i^t(y_i-\mu_i)^2|y_i\right) + \sum_i E_\theta\left(\tilde{x}_i(y_i-\mu_i)|y_i\right)E_\theta\left(\tilde{x}_i^t(y_i-\mu_i)|y_i\right) \\ &\approx \sum_i\sum_k \tilde{x}_{ik}\tilde{x}_{ik}^t V_{ik}w_{ik} \\ &\quad -\sum_i\sum_k \tilde{x}_{ik}\tilde{x}_{ik}^t(y_i-\mu_{ik})^2 w_{ik} + \sum_i\sum_k\sum_l \tilde{x}_{ik}\tilde{x}_{il}^t(y_i-\mu_{ik})(y_i-\mu_{il})w_{ik}w_{il}.\end{aligned}$$

It is also worth to consider the **a priori expected information** in (7)

$$\mathcal{I}(\theta) = E_\theta(I(y;\theta)) = \mathcal{I}_c(\theta) - E_\theta(\mathcal{I}_m(y;\theta)).$$

Because of the assumed canonical link model, the matrix $\mathcal{I}_c(y;\theta) = E_\theta(I_c(y,z;\theta)|y) = \tilde{X}^t V \tilde{X}$ is not a function of the observed data. Therefore,

$$\mathcal{I}_c(\theta) = \mathcal{I}_c(y;\theta) = \sum_i E_\theta(\tilde{x}_i \tilde{x}_i^t V_i | y_i) = \sum_i E_\theta(\tilde{x}_i \tilde{x}_i^t V_i).$$

Moreover, this also holds for

$$\begin{aligned} E_\theta\left(\sum_i E_\theta\left(\tilde{x}_i \tilde{x}_i^t (y_i - \mu_i)^2 | y_i\right)\right) &= \sum_i \int\int \tilde{x}_i \tilde{x}_i^t (y_i - \mu_i)^2 \frac{f(y_i|z_i;\theta) f(z_i)}{f(y_i;\theta)} f(y_i;\theta)\ dy\ dz \\ &= \sum_i \int \tilde{x}_i \tilde{x}_i^t \left(\int (y_i - \mu_i)^2 f(y_i|z_i;\theta)\ dy\right) f(z_i)\ dz \\ &= \sum_i \int \tilde{x}_i \tilde{x}_i^t V_i f(z_i)\ dz = \sum_i E_\theta(\tilde{x}_i \tilde{x}_i^t V_i). \end{aligned}$$

Together this gives

$$\begin{aligned} \mathcal{I}(\theta) &= E_\theta\left(\sum_i E_\theta(\tilde{x}_i (y_i - \mu_i)|y_i) E_\theta(\tilde{x}_i^t (y_i - \mu_i)|y_i)\right) = \sum_i E_\theta(s(y_i;\theta) s^t(y_i;\theta)) \\ &= \sum_i Var_\theta(s(y_i;\theta)) \qquad (10) \end{aligned}$$

the variance-covariance matrix of the observed total score. For i.i.d. variates Meilijson (1989) uses such an empirical Fisher information. Here we like to suggest to estimate the variance contribution of each individual score in an similar empirical manner for non-i.i.d. scores. Estimating the individual variance by 'samples of size one' at a time, e.g. $\widehat{Var}(s(y_i;\theta)) = s(y_i;\hat{\theta}) s^t(y_i;\hat{\theta})$, results in $\hat{\mathcal{I}} = \sum_i s(y_i;\hat{\theta}) s^t(y_i;\hat{\theta})$ with respective approximation

$$\hat{\mathcal{I}} \approx \sum_i \sum_k \sum_l \tilde{x}_{ik} \tilde{x}_{il}^t (y_i - \hat{\mu}_{ik})(y_i - \hat{\mu}_{il}) \hat{w}_{ik} \hat{w}_{il}.$$

2.2 Unspecified Random Effect Distribution

Let us again consider the model $g(\mu_i) = x_i^t\beta + z_i$. Like in the previous part, $\mu_i = E(y_i|z_i)$ denotes the conditional mean but z_i are now i.i.d. from an **unknown distribution**, which should be estimated nonparametrically. This can be done by an estimate, that is defined by giving masses $\pi = (\pi_1, \ldots, \pi_K)^t$ on a finite number K of masspoints $\zeta = (\zeta_1, \ldots, \zeta_K)^t$; both vectors are treated as unknown, whereas the number K itself is assumed to be **a priori known**. In the following the explanatory vector x should not include an intercept term! Hence, we rewrite the model as

$$g(\mu_i) = x_i^t\beta + \epsilon_i^t\zeta = \tilde{x}_i^t\theta$$

where $\tilde{x}_i = (x_i^t, \epsilon_i^t)^t$, $\theta = (\beta^t, \zeta^t)^t$ and $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{iK})^t \overset{iid}{\sim} MN(1;\pi)$, i.e. multinomial with

$$P(z_i = \zeta_k; \pi) = P(\epsilon_i = e_i; \pi) = \prod_{k=1}^{K} \pi_k^{e_{ik}}, \quad \text{with} \quad \sum_{k=1}^{K} \pi_k = 1, \quad \pi_k > 0,$$

where $e_i = (e_{i1}, \ldots, e_{iK})^t$ is any admissible realization of ϵ_i. Hence, the unknown parameters can be separated into two distinct sets of parameters, where θ corresponds to $f(y|z)$ (again a member of the exponential family) and π only belongs to the discrete estimate of $f(z)$ (the random effect density).

Let $\tilde{X} = (x_i^t, e_i^t)^t$ denote the matrix built up by all rows of the design matrix X extended by the rows of the respective row vectors e_i. The complete θ score and its negative derivative under this model is as before

$$\begin{aligned} S_c^\theta(y, z; \theta) &= \frac{\partial}{\partial\theta} \log \Big(f(y|z;\theta) f(z;\pi) \Big) = \frac{f'(y|z;\theta)}{f(y|z;\theta)} = \frac{\partial}{\partial\theta} \log f(y|z;\theta) = \tilde{X}^t(y - \mu), \\ I_c^\theta(y, z; \theta) &= -\frac{\partial^2}{\partial\theta\partial\theta^t} \log f(y|z;\theta) = \tilde{X}^t V \tilde{X}. \end{aligned}$$

To determine the π score under the multinomial model subject to the above constraint, we note that $\log f(z;\pi)$ is, aside from constants,

$$\sum_{i=1}^{n} \left(\sum_{k=1}^{K-1} e_{ik} \log \pi_k + e_{iK} \log(1 - \pi_1 - \ldots - \pi_{K-1}) \right).$$

Therefore, the complete π-score vector is

$$S_c^\pi(y, z; \pi) = \frac{\partial}{\partial\pi} \log \Big(f(y|z;\theta) f(z;\pi) \Big) = \frac{f'(z;\pi)}{f(z;\pi)} = \frac{\partial}{\partial\pi} \log f(z;\pi),$$

and equals its marginal analogue. Hence, for $k = 1, \ldots, K-1$ we get $S_c^{\pi_k}(y, z; \pi) = \sum_{i=1}^{n} \left(\frac{e_{ik}}{\pi_k} - \frac{e_{iK}}{\pi_K} \right)$ giving the vector

$$S_c^\pi(y, z; \pi) = \sum_{i=1}^{n} \sum_{k=1}^{K-1} e_k \left(\frac{e_{ik}}{\pi_k} - \frac{e_{iK}}{\pi_K} \right),$$

where $e_k = (e_{k1}, \ldots, e_{k,K-1})^t$ is a $K-1$ indicator vector with $e_{kk} = 1$ and zeros otherwise. Equating the complete score to zero results in the ML estimates

$$\hat{\pi}_c = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K-1} e_k e_{ik}.$$

Its respective negative derivative is therefore the $(K-1) \times (K-1)$ matrix

$$\begin{aligned} I_c^\pi(y, z; \pi) &= \left(-\frac{\partial}{\partial \pi_j} S_c^{\pi_k}(y, z; \pi) \right)_{k,j} = \begin{cases} \sum_{i=1}^n \left(\frac{e_{ik}}{\pi_k^2} + \frac{e_{iK}}{\pi_K^2} \right), & k = j, \\ \sum_{i=1}^n \frac{e_{iK}}{\pi_K^2}, & k \neq j. \end{cases} \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K-1} \sum_{l=1}^{K-1} e_k \frac{e_{iK}}{\pi_K^2} e_l^t + \sum_{i=1}^{n} \sum_{k=1}^{K-1} e_k \frac{e_{ik}}{\pi_k^2} e_k^t. \end{aligned}$$

Because θ and π are orthogonal parameter sets, the full matrix is

$$I_c(y, z; \theta, \pi) = \begin{pmatrix} I_c^\theta(y, z; \theta) & 0 \\ 0 & I_c^\pi(y, z; \pi) \end{pmatrix}.$$

Note that the conditional expectations can be approximated by

$$E(e_{ik}|y_i) = \sum_{j=0}^{1} j P(e_{ik} = j|y_i) = P(e_{ik} = 1|y_i) = \frac{\pi_k f(y_i|e_{ik} = 1)}{f(y_i)} \approx w_{ik}.$$

By definition, the marginal score is the conditionally expected complete score. Hence, for $k = 1, \ldots, K-1$

$$S^\pi(y; \theta, \pi) = E\Big(S_c^\pi(y, z; \pi)\Big|y\Big) = \sum_{i=1}^{n} \sum_{k=1}^{K-1} e_k E_\pi \left(\frac{e_{ik}}{\pi_k} - \frac{e_{iK}}{\pi_K} \Big| y_i \right) \approx \sum_{i=1}^{n} \sum_{k=1}^{K-1} e_k \left(\frac{w_{ik}}{\pi_k} - \frac{w_{iK}}{\pi_K} \right)$$

with respective (approximate) marginal estimates

$$\hat{\pi} = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K-1} e_k \hat{w}_{ik}.$$

The observed information is again described by the difference between the complete and the missing part. The observed θ information matrix part is handled in the same way as before. Because the θ score does not depend on π (and vice versa) the complete (θ, π) information matrix $\mathcal{I}_c(y; \theta, \pi)$ consists of two blocks. The π block is calculated by means of $\mathcal{I}_c^\pi(y; \pi)$, and $\mathcal{I}_m^\pi(y; \pi)$, the conditional variance of the complete π scores. Now we have the **complete Fisher information**

$$\begin{aligned} \mathcal{I}_c^\pi(y; \pi) &= E\Big(I_c^\pi(y, z; \pi)\Big|y\Big) = \sum_{i=1}^{n} \sum_{k=1}^{K-1} \sum_{l=1}^{K-1} e_k E\left(\frac{e_{iK}}{\pi_K^2} \Big| y_i \right) e_l^t + \sum_{i=1}^{n} \sum_{k=1}^{K-1} e_k E\left(\frac{e_{ik}}{\pi_k^2} \Big| y_i \right) e_k^t \\ &\approx \sum_{i=1}^{n} \sum_{k=1}^{K-1} \sum_{l=1}^{K-1} e_k \frac{w_{iK}}{\pi_K^2} e_l^t + \sum_{i=1}^{n} \sum_{k=1}^{K-1} e_k \frac{w_{ik}}{\pi_k^2} e_k^t \end{aligned}$$

and the **missing information**

$$\begin{aligned}\mathcal{I}_m^{\pi}(y;\pi) &= \sum_{i=1}^{n}\sum_{k=1}^{K-1}\sum_{l=1}^{K-1} e_k E\left(\Big(\frac{e_{ik}}{\pi_k}-\frac{e_{iK}}{\pi_K}\Big)\Big(\frac{e_{il}}{\pi_l}-\frac{e_{iK}}{\pi_K}\Big)\Big|y_i\right) e_l^t \\ &\quad -\sum_{i=1}^{n}\sum_{k=1}^{K-1}\sum_{l=1}^{K-1} e_k E\left(\Big(\frac{e_{ik}}{\pi_k}-\frac{e_{iK}}{\pi_K}\Big)\Big|y_i\right) E\left(\Big(\frac{e_{il}}{\pi_l}-\frac{e_{iK}}{\pi_K}\Big)\Big|y_i\right) e_l^t.\end{aligned}$$

Notice that $E(e_{ik}e_{il}|y_i) = 0$ for $k \neq l$ and $E(e_{ik}^2|y_i) = E(e_{ik}|y_i)$. Hence, the missing information equals

$$\begin{aligned}\mathcal{I}_m^{\pi}(y;\pi) &= \sum_{i=1}^{n}\sum_{k=1}^{K-1} e_k E\left(\frac{e_{ik}}{\pi_k^2}\Big|y_i\right) e_k^t + \sum_{i=1}^{n}\sum_{k=1}^{K-1}\sum_{l=1}^{K-1} e_k E\left(\frac{e_{iK}}{\pi_K^2}\Big|y_i\right) e_l^t \\ &\quad -\sum_{i=1}^{n}\sum_{k=1}^{K-1}\sum_{l=1}^{K-1} e_k E\left(\Big(\frac{e_{ik}}{\pi_k}-\frac{e_{iK}}{\pi_K}\Big)\Big|y_i\right) E\left(\Big(\frac{e_{il}}{\pi_l}-\frac{e_{iK}}{\pi_K}\Big)\Big|y_i\right) e_l^t \\ &\approx \sum_{i=1}^{n}\sum_{k=1}^{K-1} e_k \frac{w_{ik}}{\pi_k^2} e_k^t + \sum_{i=1}^{n}\sum_{k=1}^{K-1}\sum_{l=1}^{K-1} e_k \frac{w_{iK}}{\pi_K^2} e_l^t \\ &\quad -\sum_{i=1}^{n}\sum_{k=1}^{K-1}\sum_{l=1}^{K-1} e_k \Big(\frac{w_{ik}}{\pi_k}-\frac{w_{iK}}{\pi_K}\Big)\Big(\frac{w_{il}}{\pi_l}-\frac{w_{iK}}{\pi_K}\Big) e_l^t\end{aligned}$$

giving as **observed information**

$$\begin{aligned}I^{\pi}(y;\pi) &= \sum_{i=1}^{n}\sum_{k=1}^{K-1}\sum_{l=1}^{K-1} e_k E\left(\Big(\frac{e_{ik}}{\pi_k}-\frac{e_{iK}}{\pi_K}\Big)\Big|y_i\right) E\left(\Big(\frac{e_{il}}{\pi_l}-\frac{e_{iK}}{\pi_K}\Big)\Big|y_i\right) e_l^t \\ &\approx \sum_{i=1}^{n}\sum_{k=1}^{K-1}\sum_{l=1}^{K-1} e_k \Big(\frac{w_{ik}}{\pi_k}-\frac{w_{iK}}{\pi_K}\Big)\Big(\frac{w_{il}}{\pi_l}-\frac{w_{iK}}{\pi_K}\Big) e_l^t.\end{aligned}$$

3 The direct way – a reparameterized approach

We again assume that the conditional density is a member of the exponential family, i.e. for canonical link models with (extended) linear predictor $\eta_i = x_i^t\gamma + z_i$

$$f(y_i|z_i;\theta) \propto \exp\Big(y_i\eta_i - b(\eta_i)\Big) \tag{11}$$

this gives conditional moments

$$E(y_i|z_i) = \mu_i = \frac{\partial b(\eta_i)}{\partial \eta_i}$$

$$var(y_i|z_i) = V(\mu_i) = \frac{\partial^2 b(\eta_i)}{\partial \eta_i^2}$$

and first derivatives

$$\frac{\partial \log f(y_i|z_i;\theta)}{\partial \eta_i} = (y_i - \mu_i) \qquad \frac{\partial \log f(y_i|z_i;\theta)}{\partial \gamma} = (y_i - \mu_i)x_i$$
$$\frac{\partial^2 \log f(y_i|z_i;\theta)}{\partial \eta_i^2} = -V(\mu_i) \qquad \frac{\partial^2 \log f(y_i|z_i;\theta)}{\partial \gamma \partial \gamma^t} = -V(\mu_i)x_i x_i^t$$

If $f(z_i)$ does not depend on parameters, then the above results also hold for the log-likelihood based on the joint density $\log f(y_i, z_i; \theta) = \log f(y_i|z_i;\theta) + \log f(z_i)$. If $f(z_i)$ is totally unknown, we can estimate through the discrete K-point distribution given by (ζ_k, π_k), $k = 1, \ldots, K$. We first reparametrize the masses $\pi = \pi(\vartheta)$ as

$$\pi_k = \exp(\vartheta_k - \kappa(\vartheta)), \quad \text{where} \quad \partial\kappa(\vartheta)/\partial\vartheta_l = \pi_l \tag{12}$$

to ensure $\pi_k > 0$. The derivatives of π w.r.t ϑ are therefore

$$\frac{\partial \pi_k}{\partial \vartheta_l} = \begin{cases} (1-\pi_k)\pi_k & \text{if } \ k = l \\ -\pi_k \pi_l & \text{if } \ k \neq l. \end{cases} \tag{13}$$

This can be also written as

$$\frac{\partial \pi_k}{\partial \vartheta_l} = \pi_k \Big(I_{(k=l)} - \pi_l \Big).$$

Now we study the derivative of the weights

$$w_{ik} = w(y_i, \zeta_k, \gamma, \vartheta) = \frac{f(y_i|z_i;\zeta_k,\gamma)\pi_k}{\sum_{l=1}^K f(y_i|z_i;\zeta_l,\gamma)\pi_l}$$

with $f(y_i|z_i;\zeta_k,\gamma) = \exp\Big(y_i\eta_{ik} - b(\eta_{ik})\Big)$, where $\eta_{ik} = x_i^t\gamma + \zeta_k$. Therefore,

$$\begin{aligned} \frac{\partial f(y_i|z_i;\zeta_k,\gamma)}{\partial \gamma} &= x_i(y_i - \mu_{ik})f(y_i|z_i;\zeta_k,\gamma) \\ \frac{\partial f(y_i|z_i;\zeta_k,\gamma)}{\partial \zeta_l} &= I_{(k=l)}(y_i - \mu_{ik})f(y_i|z_i;\zeta_k,\gamma) \end{aligned}$$

Define $f_K(y_i) = f_K(y_i;\zeta,\gamma,\vartheta) = \sum_{k=1}^K f(y_i|z_i;\zeta_k,\gamma)\pi_k$. This gives

$$\begin{aligned} \frac{\partial f_K(y_i)}{\partial \gamma} &= \sum_{k=1}^K x_i(y_i - \mu_{ik})f(y_i|z_i;\zeta_k,\gamma)\pi_k \\ \frac{\partial f_K(y_i)}{\partial \zeta_l} &= (y_i - \mu_{il})f(y_i|z_i;\zeta_l,\gamma)\pi_l \end{aligned}$$

and

$$\frac{\partial f_K(y_i)}{\partial \vartheta_l} = \sum_{k=1}^K f(y_i|z_i;\zeta_k,\gamma)\frac{\partial \pi_k}{\partial \vartheta_l}$$

$$
\begin{aligned}
&= -\sum_{k=1}^{K} f(y_i|z_i;\zeta_k,\gamma)\pi_k\pi_l + f(y_i|z_i;\zeta_l,\gamma)\pi_l\pi_l + f(y_i|z_i;\zeta_l,\gamma)\pi_l(1-\pi_l) \\
&= -\pi_l\Big(f_K(y_i) - f(y_i|z_i;\zeta_l,\gamma)\Big)
\end{aligned}
$$

With these results we get

$$
\begin{aligned}
\frac{\partial w_{ik}}{\partial \gamma} &= \frac{1}{f_K(y_i)}\frac{\partial f(y_i|z_i;\zeta_k,\gamma)}{\partial \gamma}\pi_k - \frac{1}{f_K^2(y_i)}f(y_i|z_i;\zeta_k,\gamma)\pi_k\frac{\partial f_K(y_i)}{\partial \gamma} \\
&= \frac{x_i(y_i-\mu_{ik})f(y_i|z_i;\zeta_k,\gamma)\pi_k}{f_K(y_i)} - w_{ik}\frac{\sum_{l=1}^{K} x_i(y_i-\mu_{il})f(y_i|z_i;\zeta_l,\gamma)\pi_l}{f_K(y_i)} \\
&= x_i(y_i-\mu_{ik})w_{ik} - w_{ik}\sum_{l=1}^{K} x_i(y_i-\mu_{il})w_{il}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial w_{ik}}{\partial \zeta_l} &= \frac{1}{f_K(y_i)}\frac{\partial f(y_i|z_i;\zeta_k,\gamma)}{\partial \zeta_l}\pi_k - \frac{1}{f_K^2(y_i)}f(y_i|z_i;\zeta_k,\gamma)\pi_k\frac{\partial f_K(y_i)}{\partial \zeta_l} \\
&= \frac{I_{(k=l)}(y_i-\mu_{ik})f(y_i|z_i;\zeta_k,\gamma)\pi_k}{f_K(y_i)} - w_{ik}\frac{(y_i-\mu_{il})f(y_i|z_i;\zeta_l,\gamma)\pi_l}{f_K(y_i)} \\
&= I_{(k=l)}(y_i-\mu_{ik})w_{ik} - w_{ik}w_{il}(y_i-\mu_{il})
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial w_{ik}}{\partial \vartheta_l} &= \frac{1}{f_K(y_i)}f(y_i|z_i;\zeta_k,\gamma)\frac{\partial \pi_k}{\partial \vartheta_l} - \frac{1}{f_K^2(y_i)}f(y_i|z_i;\zeta_k,\gamma)\pi_k\frac{\partial f_K(y_i)}{\partial \vartheta_l} \\
&= \frac{1}{f_K(y_i)}f(y_i|z_i;\zeta_k,\gamma)\pi_k\Big(I_{(k=l)}-\pi_l\Big) + \frac{1}{f_K(y_i)}w_{ik}\pi_l\Big(f_K(y_i) - f(y_i|z_i;\zeta_l,\gamma)\Big) \\
&= w_{ik}(I_{(k=l)}-\pi_l) + w_{ik}\pi_l - w_{ik}w_{il} \\
&= w_{ik}\Big(I_{(k=l)} - w_{il}\Big)
\end{aligned}
$$

Let e_k be a vector of zero with 1 at the kth position. The respective derivatives of the considered estimating equations

$$
\begin{aligned}
g_\gamma(\theta,\vartheta) &= \sum_{i=1}^{n}\sum_{k=1}^{K} x_i(y_i-\mu_{ik})w_{ik} \\
g_\zeta(\theta,\vartheta) &= \sum_{i=1}^{n}\sum_{k=1}^{K} e_k(y_i-\mu_{ik})w_{ik}
\end{aligned}
$$

are

$$
\frac{\partial g_\gamma(\theta,\vartheta)}{\partial \gamma^t} = \sum_{i=1}^{n}\sum_{k=1}^{K} x_i\Big(-V_{ik}w_{ik}x_i^t + (y_i-\mu_{ik})\frac{\partial w_{ik}}{\partial \gamma^t}\Big)
$$

$$
\begin{aligned}
&= \sum_{i=1}^{n}\sum_{k=1}^{K} x_i\Big(-V_{ik}w_{ik}x_i^t + (y_i-\mu_{ik})\Big(x_i^t(y_i-\mu_{ik})w_{ik} - w_{ik}\sum_{l=1}^{K} x_i^t(y_i-\mu_{il})w_{il}\Big)\Big)\\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} x_i x_i^t w_{ik}\Big(-V_{ik}+(y_i-\mu_{ik})^2\Big) - \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{K} x_i x_i^t w_{ik}(y_i-\mu_{ik})(y_i-\mu_{il})w_{il}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial g_\gamma(\theta,\vartheta)}{\partial \zeta^t} &= \sum_{i=1}^{n}\sum_{k=1}^{K} x_i\Big(-V_{ik}w_{ik}e_k^t + (y_i-\mu_{ik})\frac{\partial w_{ik}}{\partial \zeta^t}\Big)\\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} x_i\Big(-V_{ik}w_{ik}e_k^t + (y_i-\mu_{ik})\Big(e_k^t(y_i-\mu_{ik})w_{ik} - w_{ik}\sum_{l=1}^{K} e_l^t(y_i-\mu_{il})w_{il}\Big)\Big)\\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} x_i e_k^t w_{ik}\Big(-V_{ik}+(y_i-\mu_{ik})^2\Big) - \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{K} x_i e_k^t w_{ik}(y_i-\mu_{ik})(y_i-\mu_{il})w_{il}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial g_\zeta(\theta,\vartheta)}{\partial \gamma^t} &= \sum_{i=1}^{n}\sum_{k=1}^{K} e_k\Big(-V_{ik}w_{ik}x_i^t + (y_i-\mu_{ik})\frac{\partial w_{ik}}{\partial \gamma^t}\Big)\\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} e_k\Big(-V_{ik}w_{ik}x_i^t + (y_i-\mu_{ik})\Big(x_i^t(y_i-\mu_{ik})w_{ik} - w_{ik}\sum_{l=1}^{K} x_i^t(y_i-\mu_{il})w_{il}\Big)\Big)\\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} e_k x_i^t w_{ik}\Big(-V_{ik}+(y_i-\mu_{ik})^2\Big) - \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{K} e_k x_i^t w_{ik}(y_i-\mu_{ik})(y_i-\mu_{il})w_{il}\\
&= \frac{\partial g_\gamma(\theta,\vartheta)}{\partial \zeta}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial g_\zeta(\theta,\vartheta)}{\partial \zeta^t} &= \sum_{i=1}^{n}\sum_{k=1}^{K} e_k\Big(-V_{ik}w_{ik}e_k^t + (y_i-\mu_{ik})\frac{\partial w_{ik}}{\partial \zeta^t}\Big)\\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} e_k\Big(-V_{ik}w_{ik}e_k^t + (y_i-\mu_{ik})\Big(e_k^t(y_i-\mu_{ik})w_{ik} - w_{ik}\sum_{l=1}^{K} e_l^t(y_i-\mu_{il})w_{il}\Big)\Big)\\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} e_k e_k^t w_{ik}\Big(-V_{ik}+(y_i-\mu_{ik})^2\Big) - \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{K} e_k e_l^t w_{ik}(y_i-\mu_{ik})(y_i-\mu_{il})w_{il}
\end{aligned}
$$

For $\theta = (\gamma^t, \zeta^t)^t$ and with $\tilde{x}_{ik} = (x_i^t, e_k^t)^t$ we get for the combined estimating equation

$$
g_\theta(\theta,\vartheta) = \sum_{i=1}^{n}\sum_{k=1}^{K} \tilde{x}_{ik} w_{ik}(y_i-\mu_{ik})
$$

the result

$$\frac{\partial g_\theta(\theta,\vartheta)}{\partial\theta^t} = \sum_{i=1}^{n}\sum_{k=1}^{K}\tilde{x}_{ik}\tilde{x}_{ik}^t w_{ik}\Big(-V_{ik}+(y_i-\mu_{ik})^2\Big) - \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{K}\tilde{x}_{ik}\tilde{x}_{il}^t w_{ik}(y_i-\mu_{ik})(y_i-\mu_{il})w_{il}$$

The results for the derivatives w.r.t ϑ are

$$\begin{aligned}\frac{\partial g_\gamma(\theta,\vartheta)}{\partial\vartheta^t} &= \sum_{i=1}^{n}\sum_{k=1}^{K}x_i(y_i-\mu_{ik})\frac{\partial w_{ik}}{\partial\vartheta^t}\\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}x_i(y_i-\mu_{ik})w_{ik}\Big(e_k^t-\sum_{l=1}^{K}e_l^t w_{il}\Big)\\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}x_i e_k^t w_{ik}\Big(y_i-\mu_{ik}\Big) - \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{K}x_i e_l^t w_{ik}(y_i-\mu_{ik})w_{il}\end{aligned}$$

$$\begin{aligned}\frac{\partial g_\zeta(\theta,\vartheta)}{\partial\vartheta^t} &= \sum_{i=1}^{n}\sum_{k=1}^{K}e_k(y_i-\mu_{ik})\frac{\partial w_{ik}}{\partial\vartheta^t}\\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}e_k(y_i-\mu_{ik})w_{ik}\Big(e_k^t-\sum_{l=1}^{K}e_l^t w_{il}\Big)\\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}e_k e_k^t w_{ik}\Big(y_i-\mu_{ik}\Big) - \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{K}e_k e_l^t w_{ik}(y_i-\mu_{ik})w_{il}\end{aligned}$$

Therefore,

$$\frac{\partial g_\theta(\theta,\vartheta)}{\partial\vartheta^t} = \sum_{i=1}^{n}\sum_{k=1}^{K}\tilde{x}_{ik}e_k^t w_{ik}\Big(y_i-\mu_{ik}\Big) - \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{K}\tilde{x}_{ik}e_l^t w_{ik}(y_i-\mu_{ik})w_{il}$$

The estimating equation for ϑ

$$g_\vartheta(\theta,\vartheta) = \sum_{i=1}^{K}\sum_{k=1}^{K}e_k(w_{ik}-\pi_k)$$

gives

$$\begin{aligned}\frac{\partial g_\vartheta(\theta,\vartheta)}{\partial\gamma^t} &= \sum_{i=1}^{n}\sum_{k=1}^{K}e_k\frac{\partial w_{ik}}{\partial\gamma^t}\\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}e_k x_i^t(y_i-\mu_{ik})w_{ik} - \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{K}e_k x_i^t w_{ik}(y_i-\mu_{il})w_{il}\end{aligned}$$

$$\begin{aligned}
\frac{\partial g_\vartheta(\theta,\vartheta)}{\partial \zeta^t} &= \sum_{i=1}^n \sum_{k=1}^K e_k \frac{\partial w_{ik}}{\partial \zeta^t} \\
&= \sum_{i=1}^n \sum_{k=1}^K e_k e_k^t (y_i - \mu_{ik}) w_{ik} - \sum_{i=1}^n \sum_{k=1}^K \sum_{l=1}^K e_k e_k^t w_{ik} (y_i - \mu_{il}) w_{il}
\end{aligned}$$

or in combined form

$$\frac{\partial g_\vartheta(\theta,\vartheta)}{\partial \theta^t} = \sum_{i=1}^n \sum_{k=1}^K e_k \tilde{x}_{ik}^t (y_i - \mu_{ik}) w_{ik} - \sum_{i=1}^n \sum_{k=1}^K \sum_{l=1}^K e_k \tilde{x}_{ik}^t w_{ik} (y_i - \mu_{il}) w_{il}$$

which is equal to $\partial g_\theta(\theta,\vartheta)/\partial\vartheta$.

The remaining derivative is

$$\begin{aligned}
\frac{\partial g_\vartheta(\theta,\vartheta)}{\partial \vartheta^t} &= \sum_{i=1}^n \sum_{k=1}^K e_k \Big(\frac{\partial w_{ik}}{\partial \vartheta^t} - \frac{\partial \pi_k}{\partial \vartheta^t} \Big) \\
&= \sum_{i=1}^n \sum_{k=1}^K e_k \Big(w_{ik} e_k^t - w_{ik} \sum_{l=1}^K e_l^t w_{il} - \pi_k e_k^t - \pi_k \sum_{l=1}^K e_l^t \pi_l \Big) \\
&= \sum_{i=1}^n \sum_{k=1}^K e_k e_k^t \Big(w_{ik} - \pi_k \Big) - \sum_{i=1}^n \sum_{k=1}^K \sum_{l=1}^K e_k e_l^t \Big(w_{ik} w_{il} - \pi_k \pi_l \Big).
\end{aligned}$$

Efron and Hinkley (1978)

Louis (1982)

McLachlan and Krishnan (1997)

Meilijson (1989)