A conditional likelihood approach to REML in Generalized Linear Models

Gordon K. Smyth
Department of Mathematics, University of Queensland,
Brisbane, Q 4072, Australia

Arūnas P. Verbyla
Department of Statistics, University of Adelaide,
Adelaide, SA 5005, Australia

30 May 1995

Abstract

Residual maximum likelihood estimation (REML) is often preferred to maximum likelihood estimation as a method of estimating covariance parameters in linear models because it takes account of the loss of degrees of freedom in estimating the mean and produces unbiased estimating equations for the variance parameters. In this note it is shown that REML has an exact conditional likelihood interpretation, where the conditioning is on an appropriate sufficient statistic to remove dependence on the nuisance parameters. This interpretation clarifies the motivation for REML and generalizes directly to non-normal models in which there exists a low dimensional sufficient statistic for the fitted values. The conditional likelihood is shown to be well defined and to satisfy the properties of a likelihood function, even though this is not generally true when conditioning on statistics which depend on parameters of interest. Using the conditional likelihood representation, the concept of REML is extended to generalized linear models with varying dispersion and canonical link. Explicit calculation of the conditional likelihood is given for the oneway layout. A saddle-point approximation for the conditional likelihood is also derived.

Keywords: residual maximum likelihood, restricted maximum likelihood, conditional likelihood, exponential dispersion model, modified profile likelihood, saddle-point approximation, oneway layout.

1 Introduction

Patterson and Thompson (1971) introduced residual maximum likelihood estimation (REML) as a method of estimating variance components in the context of unbalanced incomplete block designs. REML is often preferred to maximum likelihood estimation because it takes account of the loss of degrees of freedom in estimating the mean and produces unbiased estimating equations for the variance parameters. Alternative and more general derivations of REML are given by Harville (1974), Cooper & Thompson (1977) and Verbyla (1990). In all of these the residual likelihood is presented as the marginal likelihood of the error constrasts. This makes generalization of the residual likelihood principle to nonlinear models or non-normal distributions difficult since zero-mean error contrasts do not generally exist.

Cox and Reid (1987) give an approximate conditional likelihood which reduces to REML when used to estimate covariance parameters in normal linear models. Although Cox and Reid's conditional likelihood is approximate, and is based on a simplification of Barndorff-Nielsen's (1983, 1985) modified profile likelihood which reduces to REML only in special cases, it does suggest a conditional interpretation for REML. In this paper we show that REML has an exact conditional likelihood interpretation in which the conditioning is on an appropriate sufficient statistic to remove dependence on the nuisance parameters. This interpretation clarifies the motivation for REML and generalizes directly to non-normal models in which there exists a low dimensional sufficient statistic for the fitted values.

Consider the linear model $\mathbf{y} = X\boldsymbol{\beta} + \mathbf{e}$ where $\mathbf{y}$ is an $n \times 1$ vector of responses, X is an $n \times p$ design matrix of full column rank and $\mathbf{e} \sim N(0, \Omega)$ is a random vector. The covariance matrix Ω is a function of a q-dimensional parameter $\boldsymbol{\gamma}$, and is assumed positive definite for $\boldsymbol{\gamma}$ in a neighbourhood of the true value. For any fixed value of $\boldsymbol{\gamma}$, the statistic $\mathbf{t} = AX^T\Omega^{-1}\mathbf{y}$, where A is any nonsingular $p \times p$ matrix function of $\boldsymbol{\gamma}$, is complete sufficient for $\boldsymbol{\beta}$. We show that the residual likelihood can be viewed as the conditional likelihood of $\mathbf{y}$ given $\mathbf{t}$. We show, given the above form for $\mathbf{t}$, that the conditional likelihood is well defined and satisfies the properties of a likelihood function, even though this not generally true when conditioning on statistics which depend on parameters of interest.

Using the conditional likelihood representation, the concept of REML is extended to generalized linear models with varying dispersion. We assume that $y_1, \ldots, y_n$ follow a generalized linear model with canonical link, design matrix X and weights w_j/ϕ_j. The w_j are known prior weights and the ϕ_j are assumed to depend on $\boldsymbol{\gamma}$. The REML estimator of $\boldsymbol{\gamma}$ is defined to be that which maximizes the conditional likelihood of $\mathbf{y}$ given $\mathbf{t} = AX^T\Omega^{-1}\mathbf{y}$ where in this case $\Omega = \mathrm{diag}(\phi_j/w_j)$. Explicit calculation of the conditional likelihood is given for the oneway layout. A convenient saddle-point approximation is derived for use in other cases.

The idea of conditioning to remove nuisance parameters goes back at least to Bartlett (1936, 1937), and is discussed extensively by Kalbfleisch and Sprott (1970). Our conditional likelihood is direct and differs from that suggested by Kalbfleisch and Sprott and motivated by their "Euclidean assumption". The difficulties that the Euclidean assumption was intended to overcome do not occur when the conditioning statistic is of the form given above.

McCullagh and Tibshirani (1990) give an estimating equation method of adjusting profile likelihoods for nuisance parameters, which reduces to REML when estimating covariances in normal linear models. Again this is not generally equivalent to our conditional likelihood but, because it produces unbiased estimating equations, may approximate our approach in large samples.

2 Conditional Likelihood

Consider an arbitrary density or probability mass function $f(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\gamma})$ where $\boldsymbol{\beta}$ is a vector of nuisance parameters. If there exists a statistic $\mathbf{t}(\mathbf{y})$ sufficient for $\boldsymbol{\beta}$ then the nuisance parameters can be eliminated from the likelihood by conditioning on it. We have in mind cases in which f is, for fixed $\boldsymbol{\gamma}$, an exponential family so that $\mathbf{t}$ is complete sufficient and of the same dimension as $\boldsymbol{\beta}$. Let $(\mathbf{t},\mathbf{a})$ be a one-to-one transformation of $\mathbf{y}$, let $J_1 = \partial\mathbf{t}^T/\partial\mathbf{y}$ and $J_2 = \partial\mathbf{a}^T/\partial\mathbf{y}$ and let f_t be the density or probability mass function of $\mathbf{t}$. The parameter of interest, $\boldsymbol{\gamma}$, can be estimated by maximizing the conditional log-likelihood

$$\ell_{a|t}(\mathbf{y};\boldsymbol{\gamma}) = \log f(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\gamma}) - \log f_t(\mathbf{t};\boldsymbol{\beta},\boldsymbol{\gamma}) + \log|(J_1, J_2)| \tag{1}$$

which is free of $\boldsymbol{\beta}$. If the maximum likelihood estimator of $\boldsymbol{\beta}$ is a one-to-one function of $\mathbf{t}$ then it can be argued that there is no available information in $\mathbf{t}$ about $\boldsymbol{\gamma}$ in the absence of knowledge of $\boldsymbol{\beta}$, i.e., the information in $\mathbf{t}$ is entirely consumed in estimating $\boldsymbol{\beta}$. Then there should be, intuitively, no information loss in estimating $\boldsymbol{\gamma}$ from the conditional rather the full likelihood. See Sprott (1975) for an attempt to make this idea more precise.

For the above conditional likelihood to be useful it is necessary that $\mathbf{a}$ not depend on $\boldsymbol{\gamma}$; otherwise inference about $\boldsymbol{\gamma}$ would depend on the specific choice of $\mathbf{a}$. If $\mathbf{t}$ depends on $\boldsymbol{\gamma}$ it is necessary to take account of the information contained in J_1. The following lemma shows that a suitable $\mathbf{a}$ exists for the models considered in this paper and that $\mathbf{t}$ can be chosen so that the Jacobian is independent of the parameters.

Lemma. *Let $\mathbf{t} = AX^T\Omega^{-1}\mathbf{y}$ where X is an $n\times p$ design matrix of full column rank and A and Ω are full rank $p\times p$ and $n\times n$ matrices respectively depending on $\boldsymbol{\gamma}$. Let $\mathbf{a} = Z^T\mathbf{y}$ where Z is a $n\times(n-p)$ matrix of full column rank such that $X^TZ = 0$. Then $(\mathbf{t},\mathbf{a})$ is a one-to-one transformation of $\mathbf{y}$, and Jacobian of the transformation is $|Z^TZ|^{1/2}|X^TX|^{-1/2}|X^T\Omega^{-1}X||A|$.*

Proof. $\Omega^{-1/2}X$ and $\Omega^{1/2}Z$ are orthogonal and of full rank. Therefore $(\Omega^{-1/2}X, \Omega^{1/2}Z)$ is nonsingular, as is

$$\Omega^{-1/2}(\Omega^{-1/2}X, \Omega^{1/2}Z)\begin{pmatrix} A^T & 0 \\ 0 & I_{n-p}\end{pmatrix} = (\Omega^{-1}XA^T, Z).$$

This shows that $(\mathbf{t},\mathbf{a})$ is a one-to-one transformation. The Jacobian is

$$|(J_1, J_2)| = \begin{vmatrix} J_1^TJ_1 & J_1^TJ_2 \\ J_2^TJ_1 & J_2^TJ_2 \end{vmatrix}^{1/2} = |J_2^TJ_2|^{1/2}|J_1^TJ_1 - J_1^TJ_2(J_2^TJ_2)^{-1}J_2^TJ_1|^{1/2}.$$

Now $J_2 = Z$ and $I - Z(Z^T Z)^{-1} Z^T$ is the orthogonal projection onto the null space of Z and is therefore equal to $X(X^T X)^{-1} X^T$, so

$$|(J_1, J_2)| = |Z^T Z|^{1/2} |J_1^T X (X^T X)^{-1} X^T J_1|^{1/2}.$$

Putting $J_1 = \Omega^{-1} X A$ and factoring the determinant gives the required result. $\square$

The conditional likelihood (1) is invariant with respect to A. A convenient choice is $A = (X^T \Omega^{-1} X)^{-1}$ because then the Jacobian becomes $|Z^T Z|^{1/2} |X^T X|^{1/2}$ which is independent of the parameters. Without loss of generality we choose $|Z^T Z| = 1$ so that $\mathbf{a}$ is a volume-preserving function of $\mathbf{y}$. We define the conditional log-likelihood of $\mathbf{y}$ given $\mathbf{t}$ to be

$$\ell_{y|t}(\mathbf{y}; \boldsymbol{\gamma}) = \log f(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\gamma}) - \log f_{\tilde{\beta}}(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\gamma}) + \frac{1}{2} \log |X^T X|$$

where $f_{\tilde{\beta}}$ is the density or probability mass function of $\tilde{\boldsymbol{\beta}} = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} \mathbf{y}$. Since f and $f_{\tilde{\beta}}$ integrate to one for all $\boldsymbol{\gamma}$, $\ell_{y|t}$ satisfies the usual properties of a log-likelihood function in that $\dot{\ell}_{y|t} = \partial \ell_{y|t} / \partial \boldsymbol{\gamma}$ has expectation zero and $\mathrm{var}(\dot{\ell}_{y|t}) = E(-\partial^2 \ell_{y|t} / \partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T)$.

Consider now the linear model described in the introduction. The residual likelihood for $\boldsymbol{\gamma}$ is usually defined to be the marginal likelihood of $\mathbf{a} = Z^T \mathbf{y}$ where Z is as above. Since $\mathbf{a}$ and $\mathbf{t}$ are in this case independent, it is trivially true that the residual likelihood is the conditional likelihood of $\mathbf{y}$ given $\mathbf{t}$. Since $\tilde{\boldsymbol{\beta}} \sim N\{\boldsymbol{\beta}, (X^T \Omega^{-1} X)^{-1}\}$, we calculate

$$\begin{aligned} \ell_{y|t}(\mathbf{y}; \boldsymbol{\gamma}) &= -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\Omega| - \frac{1}{2} (\mathbf{y} - X\boldsymbol{\beta})^T \Omega^{-1} (\mathbf{y} - X\boldsymbol{\beta}) + \frac{p}{2} \log(2\pi) \\ &\quad - \frac{1}{2} \log |X^T \Omega^{-1} X| + \frac{1}{2} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T X^T \Omega^{-1} X (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \frac{1}{2} \log |X^T X| \\ &= \frac{n-p}{2} \log(2\pi) - \frac{1}{2} \log |\Omega| - \frac{1}{2} \log |X^T \Omega^{-1} X| - \frac{1}{2} \mathbf{y}^T P \mathbf{y} + \frac{1}{2} \log |X^T X| \end{aligned}$$

where $P = \Omega^{-1} - \Omega^{-1} X (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1}$. This is identical to the residual likelihood function given by Harville (1974) and Cooper and Thompson (1977).

3 Generalized Linear Models

Consider the probability density function defined by

$$f(y; \theta, \phi) = \exp[\{y\theta - \kappa(\theta)\}/\phi + c(y, \phi)] \tag{2}$$

where $\phi > 0$ and $\theta \in \Theta = \{\theta' : \kappa(\theta') < \infty\}$. Following Jørgensen (1987), the distribution defined by $f(y; \theta, \phi)$ is called an exponential dispersion model with dispersion parameter ϕ, and is denoted $\mathrm{ED}(\mu, \phi)$ where $\mu = E(y) = \dot{\kappa}(\theta)$. The cumulant function $\kappa()$ can always be chosen so that $\kappa(0) = 0$ and $\exp c(y, \phi) = f(y; 0, \phi)$, and assuming this has been done let $\mathrm{C}(\phi)$ denote the distribution with this latter density. In that case $s \to \kappa(s)/\phi$ is the cumulant generating function of

y/ϕ with $y \sim \mathrm{C}(\phi)$ and $s \to \{\kappa(s+\theta)-\kappa(\theta)\}/\phi$ is the cumulant generating function of y/ϕ with $y \sim \mathrm{ED}(\mu, \phi)$. The process by which the exponential dispersion model is generated from the base density $\exp c(y, \phi)$ is often called exponential tilting.

Let $y_j \sim \mathrm{ED}(\mu_j, \phi_j/w_j)$, $j = 1, \ldots, n$, be independent random variables where the w_j are known weights. A generalized linear model arises if a link-linear predictor is assumed for the means, $g(\mu_j) = \mathbf{x}_j^T\boldsymbol{\beta}$ where $\mathbf{x}_j$ is a vector of covariates, $\boldsymbol{\beta}$ is an unknown p-vector of regression parameters and $g()$ is a known link function. We assume also that the dispersions ϕ_j depend on an unknown parameter vector $\boldsymbol{\gamma}$, for example through a link-linear predictor $h(\phi_j) = \mathbf{z}_j^T\boldsymbol{\gamma}$ as in Smyth (1989), where $\mathbf{z}_j$ is a vector of covariates and $\boldsymbol{\gamma}$ is an unknown parameter vector.

Let $\Omega = \mathrm{diag}(\phi_j/w_j)$ and X be the $n \times p$ matrix with $\mathbf{x}_j^T$ as ith row. We assume $g()$ to be the canonical link function such that $g(\mu_j) = \theta_j$, so that $\mathbf{t} = X^T\Omega^{-1}\mathbf{y}$ is a complete sufficient statistic for $\boldsymbol{\beta}$. We define the REML estimate of $\boldsymbol{\gamma}$ to be that which maximizes the conditional likelihood of $\mathbf{y}$ given $\mathbf{t}$.

The conditional log-likelihood of $\mathbf{y}$ given $\mathbf{t}$ is

$$\ell_{y|t}(\mathbf{y};\boldsymbol{\gamma}) = \log f(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\gamma}) - \log f_t(\mathbf{t};\boldsymbol{\beta},\boldsymbol{\gamma}) + \frac{1}{2}\log|X^TX| - \log|X^T\Omega^{-1}X|.$$

For fixed ϕ_j the sufficient statistic $\mathbf{t}$ has itself an linear exponential family distribution with cumulant function

$$\kappa_t(\boldsymbol{\beta}) = \sum_{j=1}^n w_j\phi_j^{-1}\kappa(\mathbf{x}_j^T\boldsymbol{\beta})$$

where $\kappa()$ is the cumulant function of the y_j. The cumulant generating function of $\mathbf{t}$ is $K_t(\mathbf{s}) = \kappa_t(\boldsymbol{\beta}+\mathbf{s}) - \kappa_t(\boldsymbol{\beta})$, so the probability density function of $\mathbf{t}$ is given by

$$f_t(\mathbf{t};\boldsymbol{\beta},\boldsymbol{\gamma}) = \frac{1}{(2\pi i)^p}\int \exp\left\{\kappa_t(\boldsymbol{\beta}+\mathbf{s}) - \kappa(\boldsymbol{\beta}) - \mathbf{s}^T\mathbf{t}\right\} d\mathbf{s}$$

where the integral is taken over the imaginary axis with respect to each variable. This can be factorized

$$\begin{aligned} f_t(\mathbf{t};\boldsymbol{\beta},\boldsymbol{\gamma}) &= \exp\{\mathbf{t}^T\boldsymbol{\beta} - \kappa_t(\boldsymbol{\beta})\}\frac{1}{(2\pi i)^p}\int \exp\left\{\kappa_t(\boldsymbol{\beta}+\mathbf{s}) - \mathbf{t}^T(\boldsymbol{\beta}+\mathbf{s})\right\} d\mathbf{s} \\ &= \exp\{\mathbf{t}^T\boldsymbol{\beta} - \kappa_t(\boldsymbol{\beta})\}\frac{1}{(2\pi i)^p}\int \exp\left\{\kappa_t(\mathbf{s}) - \mathbf{t}^T\mathbf{s}\right\} d\mathbf{s} \end{aligned}$$

provided that $\kappa_t(\mathbf{s})$ has no singularities in any s_j between the imaginary axis and the imaginary axis shifted by β_j. This shows that

$$\begin{aligned} \ell_{y|t}(\mathbf{y};\boldsymbol{\gamma}) &= \sum_{j=1}^n c(y_j, \phi_j/w_j) - \log\frac{1}{(2\pi i)^p}\int \exp\left\{\kappa_t(\mathbf{s}) - \mathbf{t}^T\mathbf{s}\right\} d\mathbf{s} \\ &\quad - \log|X^T\Omega^{-1}X| + \frac{1}{2}\log|X^TX| \end{aligned} \tag{3}$$

which exhibits lack of dependence on $\boldsymbol{\beta}$.

Now $\kappa_t(\mathbf{s})$ would be the cumulant generating function of $\mathbf{t}$ if the y_j were distributed according to the base distributions $\mathrm{C}(\phi_j/w_j)$. Therefore (3) can be viewed as the conditional likelihood of $\mathbf{y}$ given $\mathbf{t}$ with the y_j sampled from $\mathrm{C}(\phi_j/w_j)$ rather than from $\mathrm{ED}(\mu_j, \phi_j/w_j)$. In the full density (2) the exponential family parameter θ determines the location of the distribution while ϕ determines the dispersion. Intuitively, the effect of conditioning on the sufficient statistic $\mathbf{t}$ is to reverse the process of exponential tilting, returning to the base distribution defined by $c(y, \phi)$ in which only the dispersion parameter appears.

In many cases it will be inconvenient to compute the integral in (3). The saddle-point approximation to the density which the integral represents is

$$(2\pi)^{-p/2}|\mathcal{I}(\hat{\mathbf{s}})|^{-1/2}\exp\{\kappa_t(\hat{\mathbf{s}}) - \mathbf{t}^T\hat{\mathbf{s}}\}$$

where $\hat{\mathbf{s}}$ solves $\dot{\kappa}_t(\mathbf{s}) = \mathbf{t}$ and $\mathcal{I}(\mathbf{s}) = \ddot{\kappa}_t(\mathbf{s})$. Now $\mathbf{t} - \dot{\kappa}_t(\mathbf{s}) = X^T\Omega^{-1}\{\mathbf{y} - \boldsymbol{\mu}(\mathbf{s})\}$, so $\hat{\mathbf{s}}$ solves the normal equations for the generalized linear model and is therefore the maximum likelihood estimator $\hat{\mathbf{s}} = \hat{\boldsymbol{\beta}}_\gamma$ for $\boldsymbol{\beta}$ with $\boldsymbol{\gamma}$ fixed. Also $\ddot{\kappa}_t(\hat{\mathbf{s}}) = X^TWX$, where $W = \mathrm{diag}\{V(\hat{\mu}_j)w_j/\phi_j\}$, which is the estimated information matrix for β with $\boldsymbol{\gamma}$ fixed. The saddle-point approximation therefore is

$$(2\pi)^{-p/2}|X^TWX|^{-1/2}\exp\{\kappa_t(\hat{\boldsymbol{\beta}}_\gamma) - \mathbf{t}^T\hat{\boldsymbol{\beta}}_\gamma\}$$

and the approximate conditional likelihood is

$$\ell_{y|t}(\mathbf{y};\boldsymbol{\gamma}) \approx \log f(\mathbf{y};\hat{\boldsymbol{\beta}}_\gamma,\boldsymbol{\gamma}) + \frac{p}{2}\log 2\pi + \frac{1}{2}\log|X^TWX| - \log|X^T\Omega^{-1}X| + \frac{1}{2}\log|X^TX|.$$

This reduces to the usual REML likelihood for normal data, when $W = \Omega^{-1}$. It is equivalent to the approximate conditional likelihood of Cox and Reid (1987) when the estimating a constant dispersion. In other cases it differs from Cox and Reid's approximate conditional likelihood in that the information matrix $(X^TWX)^{-1}$ is replaced by $(X^T\Omega^{-1}X)^{-1}X^TWX(X^T\Omega^{-1}X)^{-1}$.

The final section of this paper works out REML estimators for certain generalized linear models in which the conditional likelihood can be obtained in closed form.

4 The Oneway Layout

Consider a generalized linear model with means described by a one-way classification, i.e., let y_{jk}, $j = 1, \ldots, b$, $k = 1, \ldots, n_j$, be independent random variables with $y_{jk} \sim \mathrm{ED}(\beta_j, \phi_{jk})$ where the ϕ_{jk} are functions of a q-vector of parameters $\boldsymbol{\gamma}$. Such a model arises where there is only one factor in an experiment or when all interactions are being estimated in a multi-factor experiment.

Consider first the case in which the dispersions are constant, $\phi_{jk} = \gamma$ for all j and k. The group mean $\bar{y}_j$ is sufficient for β_j and is distributed as $\mathrm{ED}(\beta_j, \gamma/n_j)$. The conditional log-likelihood is

$$\begin{aligned}\ell_{y|\hat{\beta}}(\mathbf{y};\gamma) &= \sum_{j=1}^{b}\left\{\sum_{k=1}^{n_j}\log f(y_{jk};\theta_j,\gamma) - \log f(\bar{y}_j;\theta_j,\gamma/n_j) + \frac{1}{2}\log n_j\right\}\\ &= \sum_{j=1}^{b}\left\{\sum_{k=1}^{n_j}c(y_{jk},\gamma) - c(\bar{\mathbf{y}}_j,\gamma/n_j) + \frac{1}{2}\log n_j\right\}.\end{aligned}$$

In the normal and inverse-Gaussian cases the REML estimator of γ is the residual mean deviance. In these cases the estimator $\hat{\gamma}$ is uniform minimum variance unbiased for γ, and $(N-b)\hat{\gamma}/\gamma \sim \chi^2_{N-b}$ independently of the $\bar{y}_j$.

For the gamma distribution we can take $c(y,\gamma) = \log(y/\gamma)/\gamma - \log\Gamma(1/\gamma) - \log y$ so

$$
\begin{aligned}
\ell_{y|\hat{\beta}} &= \frac{1}{\gamma}\sum_{j=1}^{b}\sum_{k=1}^{n_j}\log(y_{jk}/\bar{y}_j) - N\log\Gamma(1/\gamma) \\
&\quad + \sum_{j=1}^{b}\log\Gamma(n_j/\gamma) - \sum_{j=1}^{b}\left(\sum_{k=1}^{n_j}\log y_{jk} - \log\bar{y}_j\right).
\end{aligned}
$$

This is an exponential family likelihood with canonical parameter $\nu = 1/\gamma$, sufficient statistic $D(\mathbf{y}) = \sum_{j=1}^{b}\sum_{k=1}^{n_j}\log(y_{jk}/\bar{y}_j)$ and cumulant function $\lambda(\nu) = N\log\Gamma(\nu) - \sum_{j=1}^{b}\log\Gamma(n_j\nu)$. The REML estimator of γ is obtained by equating $D(\mathbf{y})$ to its expectation,

$$
D(\mathbf{y}) = \dot{\lambda}(\nu) = N\psi(\nu) - \sum_{j=1}^{b} n_j\psi(n_j\nu)
$$

where $\psi()$ is the digamma function. This can be compared to maximum likelihood estimation of γ which would have $\log(\nu)$ in place of $\psi(n_j\nu)$ in the last term. Compare with Cox and Reid (1987, p. 12) and McCullagh and Nelder (1989, p. 295).

Now consider the general case in which the ϕ_{jk} are general functions of $\boldsymbol{\gamma}$. The log-likelihood is

$$
\begin{aligned}
\ell_y &= \sum_{j=1}^{b}\sum_{k=1}^{n_j}\left[\frac{1}{\phi_{jk}}\{y_{jk}\theta_j - \kappa(\theta_j)\} + c(y_{jk},\phi_{jk})\right] \\
&= \sum_{j=1}^{b}\left[\frac{1}{\alpha_j}\{t_j\theta_j - \kappa(\theta_j)\} + \sum_{k=1}^{n_j} c(y_{jk},\phi_{jk})\right]
\end{aligned}
$$

where $\alpha_j = (\sum_{k=1}^{n_j}\phi_{jk}^{-1})^{-1}$, $t_j = \alpha_j\sum_{k=1}^{n_j}\phi_{jk}^{-1}y_{jk}$ and $\beta_j = \dot{\kappa}(\theta_j)$. Each t_j is sufficient for β_j and is distributed as $\mathrm{ED}(\beta_j,\alpha_j)$. The conditional log-likelihood of $\mathbf{y}$ given the t_j is

$$
\ell_{y|t} = \sum_{j=1}^{b}\left\{\sum_{k=1}^{n_j} c(y_{jk},\phi_{jk}) - c(t_j,\alpha_j)\right\}.
$$

If the y_{jk} are normally distributed, this is equivalent to the usual REML log-likelihood given in Section 2. If the y_{jk} are gamma, the function $c(y,\gamma)$ is as given above. If the y_{jk} are inverse-Gaussian, we can take $c(y,\gamma) = 1/(2\gamma y) - (1/2)\log\gamma - (3/2)\log y - (1/2)\log 2\pi$.

Acknowledgements

The authors are grateful for discussions with Professor Alan James and thank three anonymous referees for suggestions which led to improvements in the paper.

References

Barndorff-Nielsen, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, **70**, 343–365.

Barndorff-Nielsen, O. E. (1985). Properties of modified profile likelihood. In *Contributions to Probability and Statistics in Honour of Gunnar Blom* (J. Lanke and G. Lindgren, eds), pp. 25–38. Lund.

Bartlett, M. S. (1936). The information available in small samples. *Proc. Camb. Phil. Soc.*, **32**, 560–566.

Bartlett, M. S. (1937) Properties of sufficiency and statistical tests. *Proc. R. Soc.* A, **160**, 268–282.

Cooper, D. M. and Thompson, R. (1977). A note on the estimation of the parameters of the autoregressive-moving average process. *Biometrika*, **64**, 625–628.

Cox, D. R. and Reid, N. (1987) Parameter orthogonality and approximate conditional inference. *J. R. Statist. Soc.* B, **49**, 1–39.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, **61**, 383–385.

Jørgensen, B. (1987). Exponential dispersion models (with discussion). *J. R. Statist. Soc. B*, **49**, 127–162.

Kalbfleisch, J. D. and Sprott, D. A. (1970). Application of likelihood methods to models involving a large number of nuisance parameters. *J. R. Statist. Soc.* B, **32**, 175–208.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models, 2nd ed.* Chapman and Hall: London.

McCullagh, P. and Tibshirani, R. (1990). A simple method for the adjustment of profile likelihoods. *J. R. Statist. Soc.* B, **52**, 325–344.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–554.

Smyth, G. K. (1989). Generalized linear models with varying dispersion. *J. Roy. Statist. Soc.* B **51**, 47–60.

Sprott, D. A. (1975). Marginal and conditional sufficiency. *Biometrika*, **62**, 599–605.

Verbyla, A. P. (1990). A conditional derivation of residual maximum likelihood. *Aust. J. Statist.*, **32**, 221–224.