Bias Reduction in Exponential Family Nonlinear Models

Ioannis Kosmidis and David Firth

CRiSM and Department of Statistics, University of Warwick
Coventry CV4 7AL, UK

October 16, 2008

Abstract

In Firth (1993, *Biometrika*) it was shown how the leading term in the asymptotic bias of the maximum likelihood estimator is removed by adjusting the score vector, and that in canonical-link generalized linear models the method is equivalent to maximizing a penalized likelihood which is easily implemented via iterative adjustment of the data. Here a more general family of bias-reducing adjustments is developed, for a broad class of univariate and multivariate generalized nonlinear models. The resulting formulae for the adjusted score vector are computationally convenient, and in univariate models they directly suggest implementation through an iterative scheme of data adjustment. For generalized linear models a necessary and sufficient condition is given for the existence of a penalized likelihood interpretation of the method. An illustrative application to the Goodman row-column association model shows how the computational simplicity and statistical benefits of bias reduction extend beyond generalized linear models.

1 Introduction

In regular parametric statistical models the maximum likelihood estimator is consistent, and the leading term in its asymptotic bias expansion is of magnitude $O(n^{-1})$; here n denotes the sample size or other index of information, assumed to be large relative to the number, p say, of parameters. Among methods that have been suggested for removal of the $O(n^{-1})$ bias, the approach taken in Firth (1993) has received considerable recent attention. That method operates by making a suitably constructed adjustment to the score vector, and has the advantage of not requiring the value of the maximum likelihood estimate itself. This last fact has, at least in part, motivated detailed empirical studies of the method in some common situations where the maximum likelihood estimate can be infinite, notably logistic or similar regression models for a binary or multinomial response, and models for censored lifetime data; see, for example, Mehrabi & Matthews (1995), Pettitt et al. (1998), Heinze & Schemper (2002), Bull et al. (2002), Heinze & Schemper (2004), Zorn (2005), Sartori (2006) and Bull et al. (2007).

Point estimation and unbiasedness are, of course, not strong statistical principles. The notion of bias, in particular, relates to a specific parameterization of a model: for example, unbiasedness of the familiar sample variance S^2 as an estimator of σ^2 does not deliver an unbiased estimator of σ itself. Thus bias correction of the maximum likelihood estimator, by any method, violates the appealing property of exact equivariance under reparameterization of the model. Moreover, as noted by Firth (1993) and well known previously, reduction in bias

may sometimes be accompanied by inflation of variance, possibly yielding an estimator whose mean squared error is worse than that of maximum likelihood. Despite these reservations, it is amply evident from published empirical studies such as those mentioned above that, in some very commonly used models, bias reduction by the method studied here can perform substantially better than unadjusted maximum likelihood, especially when the sample size is not large. Importantly, 'better' here means not only in terms of bias, but also other properties such as finiteness, mean squared error, and the coverage of approximate confidence intervals. The remarks above make plain that such improvements cannot be universal. A primary motivation for the present paper is to provide a unified conceptual and computational framework for further exploration of the properties of bias-reduced maximum likelihood in a large class of models which includes many model types commonly used in practice. The class of multivariate generalized nonlinear models includes univariate Gaussian and generalized nonlinear regressions, and univariate and multivariate generalized linear models, as prominent special cases.

The development will be broadly from the general to the specific. In Section 2 the method of Firth (1993) is embedded within an extended family of bias-reducing adjustments to the score vector. Section 3 then gives the central results of the paper, namely the unified derivation of score adjustments for the broad class of multivariate generalized nonlinear models with known dispersion, including as a special case the multivariate generalized linear model. Section 4 exploits this general development to give new results for univariate generalized linear models with non-canonical link, specifically a necessary and sufficient condition for the existence of a penalized likelihood interpretation and the corresponding family of penalized likelihoods, as well as the detail of an iterated data-adjustment approach to computation; the latter connects with earlier results of Cordeiro & McCullagh (1991). In Section 5 the details of implementation are extended to the special case of a univariate generalized nonlinear model; these results are then applied in Section 6 to a specific example, the RC(1) association model for the analysis of a two-way contingency table (Goodman, 1979, 1985). The results of a small simulation study in Section 6 suggest that such row-column multiplicative association models represent a further promising application context, additional to those already highlighted in the published work mentioned above and elsewhere, in which bias reduction of maximum likelihood estimates is likely to be routinely beneficial.

2 Bias reduction by adjustment of the score

2.1 Preamble

In regular problems, maximum likelihood estimates are obtained by solving the score equation $U(\beta) = \nabla l(\beta) = 0$, where $l(\beta)$ is the log-likelihood function for parameter vector β. Under regularity conditions in the spirit of those in Cramér (1946, §33.2), which guarantee the consistency of the maximum likelihood estimator, Firth (1993) showed how, by solving a suitably adjusted version of the score equation, removal of the $O(n^{-1})$ bias term of the maximum likelihood estimator is achieved. In Firth (1993) the form of the adjusted score vector is given in index notation and using the Einstein summation convention. Despite the general elegance and compactness of such notation it will be departed from here, since matrix notation will be simpler and more illuminating for the present purposes. First some notational rules are introduced for the representation of arrays as blocked matrices.

2.2 Notation

Consider a sequence of 4-way arrays $\{E_r;\ r = 1, \ldots, k\}$. Such an array E_r is an arrangement of scalars e_{rstuv} with $s \in \{1, \ldots, l\}$, $t \in \{1, \ldots, m\}$, $u \in \{1, \ldots, n\}$ and $v \in \{1, \ldots, q\}$. The array E_r can be represented as a $ln \times mq$ blocked matrix having the form

$$E_r = \begin{bmatrix} E_{r11} & E_{r12} & \cdots & E_{r1m} \\ E_{r21} & E_{r22} & \cdots & E_{r2m} \\ \vdots & \vdots & \ddots & \vdots \\ E_{rl1} & E_{rl2} & \cdots & E_{rlm} \end{bmatrix},$$

with E_{rst} a $n \times q$ matrix. Denote by E_{rstu} the uth row of the matrix E_{rst} as a row vector, i.e. having dimension $1 \times q$.

If $m = 1$, the representation of a sequence of 3-way arrays results, and E_r has sth block the $n \times q$ matrix E_{rs}. In this case, E_{rst} denotes the tth row of E_{rs}, as a row vector, i.e. having dimension $1 \times q$. A sequence $\{E_r;\ r = 1, \ldots, k\}$ of 2-way arrays of dimension $n \times q$ results from setting $m = 1$ and $l = 1$ in the 4-way case. In this case, E_{rs} denotes the sth row of E_r, as a row vector, i.e. having dimension $1 \times q$. The blocking structure and dimensions of any quantity appearing below will be clearly stated or should be self-evident from the context.

For brevity, the dependence of array structures on β will usually be suppressed.

2.3 General family of adjustments

With the above conventions let F and I denote the Fisher and observed information on the p-dimensional parameter β. Suppose that the tth component U_t of the score vector is adjusted to

$$U_t^* = U_t + A_t \quad (t = 1, \ldots, p),$$

where A_t is $O_p(1)$ as $n \to \infty$. As in Firth (1993) but now in matrix notation, removal of the $O(n^{-1})$ bias term occurs if A_t is either

$$A_t^{(E)} = \frac{1}{2} \operatorname{tr} \left\{ F^{-1}(P_t + Q_t) \right\}$$

or

$$A_t^{(O)} = I_t F^{-1} A^{(E)},$$

based on the expected (Fisher) or observed information, respectively. In the above, $P_t = E(UU^T U_t)$ is the tth block of the $p^2 \times p$ third-order cumulant matrix of the score vector, and $Q_t = E(-IU_t)$ is the tth block of the $p^2 \times p$ blocked matrix of covariances of the first and second derivatives of the log-likelihood with respect to the parameters. The $1 \times p$ vector I_t denotes the tth row of I. All expectations are taken with respect to the model and at β. These adjustments based on observed and expected information are in fact particular instances of bias-reducing adjustments in the more general family

$$A_t = (I_t + R_t) F^{-1} A^{(E)}, \tag{1}$$

where R_t is $O_p(n^{1/2})$. The adjustments (1) can usefully be re-expressed as a weighted sum of the adjustments based on the expected information, so that the components of the adjusted score vector take the general form

$$U_t^* = U_t + \sum_{u=1}^{p} e_{tu} A_u^{(E)} \quad (t = 1, \ldots, p). \tag{2}$$

Here e_{tu} denotes the (t,u)th component of matrix $(I+R)F^{-1}$; in the special case of adjustment based on the expected information this is simply the identity matrix.

3 Adjusted score for multivariate exponential family models

3.1 The multivariate generalized nonlinear model

Consider a q-dimensional random variable Y from the exponential family having density or probability mass function of the form

$$f(y\,;\theta) = \exp\left\{\frac{y^T\theta - b(\theta)}{\lambda} + c(y,\lambda)\right\},$$

in which the dispersion λ is assumed known. The expectation and the $q\times q$ variance-covariance matrix of Y are

$$\begin{aligned} E(Y\,;\theta) &= \nabla b(\theta) = \mu(\theta)\,, \\ \mathrm{cov}\,(Y\,;\theta) &= \lambda\mathcal{D}^2\left(b(\theta);\theta\right) = \Sigma(\theta)\,. \end{aligned}$$

The quantity $\mathcal{D}^2\left(b(\theta);\theta\right)$ is the $q\times q$ Hessian matrix of $b(\theta)$ with respect to θ. Generally, in what follows, if a and b are p and q dimensional vectors, respectively, $\mathcal{D}\left(a;b\right)$ is the $p\times q$ matrix of first derivatives of a with respect to b, and $\mathcal{D}^2\left(a;b\right)$ is a $pq\times q$ blocked Hessian matrix of the second derivatives of a with respect to b, having blocks the $q\times q$ matrices $\mathcal{D}^2\left(a_i;b\right)$ $(i=1,\ldots,p)$.

Suppose now that observed response vectors $y_1,\ldots,y_n$ are realizations of independent random vectors $Y_1,\ldots,Y_n$, with θ_r the parameter vector for Y_r $(r=1,\ldots,n)$ and λ_r the known dispersion. An exponential family nonlinear model, or generalized nonlinear model, connects the expectation of each Y_r with a predictor η_r through a known vector-valued link function $g_r:\Re^q\to\Re^q$,

$$g_r(\mu_r) = \eta_r(\beta) \quad (r=1,\ldots,n)\,, \tag{3}$$

where $\eta_r:\Re^p\to\Re^q$ is a specified vector-valued function of model parameters β that is at least twice continuously differentiable. For technical details on the required regularity conditions and the properties of this class of models see Wei (1997).

With the notation $\mu_r=\mu(\theta_r(\beta))$ and $\Sigma_r=\Sigma(\theta_r(\beta))$, the score vector for the parameters β of model (3) has the form

$$U = \sum_r X_r^T W_r \mathcal{D}\left(\eta_r;\mu_r\right)\left(y_r-\mu_r\right), \tag{4}$$

where $W_r = D_r\Sigma_r^{-1}D_r^T$ is the generalization of the *working weight*, defined exactly as for generalized linear multivariate models (Fahrmeir & Tutz, 2001, Appendix A.1). Also, $X_r = \mathcal{D}\left(\eta_r;\beta\right)$, and $D_r^T = \mathcal{D}\left(\mu_r;\eta_r\right)$. The observed information on β is given by

$$\begin{aligned} I = \sum_r X_r^T W_r X_r - \sum_r\sum_{s=1}^{q}\lambda_r^{-1}\left(X_r^T V_{rs} X_r\right)\left(y_{rs}-\mu_{rs}\right) \\ - \sum_r\sum_{s,u=1}^{q}\mathcal{D}^2\left(\eta_{ru};\beta\right)\left(\Sigma_r^{-1}D_r^T\right)_{su}\left(y_{rs}-\mu_{rs}\right), \end{aligned} \tag{5}$$

where $V_{rs}=\mathcal{D}^2\left(\theta_{rs};\eta_r\right)$ and $(\Sigma_r^{-1}D_r^T)_{su}$ is the (s,u)th element of the product $\Sigma_r^{-1}D_r^T$. The Fisher information is $F=E(I)=\sum_r X_r^T W_r X_r$.

3.2 Derivation of the adjusted score

From (4) and (5) and after some algebra, the sum of the cumulant matrices P_t and Q_t is found to be

$$P_t + Q_t = \sum_r \sum_{s=1}^{q} [X_r^T \left\{ \left(D_r \Sigma_r^{-1} \right)_s \otimes 1_q \right\} \mathcal{D}^2 \left(\mu_r ; \eta_r \right) X_r + (W_{rs} \otimes 1_q) \mathcal{D}^2 \left(\eta_r ; \beta \right)] x_{rst} \,, \tag{6}$$

where W_{rs} is the sth row of the $q \times q$ matrix W_r as a $1 \times q$ vector, and $(D_r \Sigma_r^{-1})_s$ is the sth row of $D_r \Sigma_r^{-1}$ as a $1 \times q$ vector.

After substituting (6) into (2), and some tedious but straightforward algebraic manipulation, the adjusted score components for exponential family nonlinear models with known dispersion are found to take the form

$$\begin{aligned} U_t^* = U_t + \frac{1}{2} \sum_r \sum_{s=1}^{q} \operatorname{tr} \big[H_r W_r^{-1} \left\{ \left(D_r \Sigma_r^{-1} \right)_s \otimes 1_q \right\} \mathcal{D}^2 \left(\mu_r ; \eta_r \right) \\ + F^{-1} (W_{rs} \otimes 1_q) \mathcal{D}^2 \left(\eta_r ; \beta \right) \big] x_{rst}^* \quad (t = 1, \ldots, p) \,, \end{aligned} \tag{7}$$

where $x_{rst}^* = \sum_{u=1}^{p} e_{tu} x_{rsu}$, and $H_r = X_r F^{-1} X_r^T W_r$ $(r = 1, \ldots, n)$ are the $q \times q$ diagonal blocks of the projection or 'hat' matrix which have the well-known leverage interpretation in linear models. For generalized linear models, H_r can still be used as a generalized leverage measure, but its key characteristic in the context of the present work is that it projects the current 'working observation' for the rth subject to the next value of the linear predictor in a Fisher scoring iteration. For a description of the generalized leverage and how it can be used, see Fahrmeir & Tutz (2001, §4.2.2).

Expression (7) is a convenient representation for computational purposes, especially, as the quantities involved are all routinely available after fitting a model by maximum likelihood in a typical software package. A further appealing feature of this representation is that its structure is unchanged for different choices from the generic family of adjustments (1): only the coefficients e_{tu} change.

Consider now the special case of the canonical link function g such that $\theta_r = g(\mu_r) = \eta_r$ $(r = 1, \ldots, n)$. This definition of canonical link extends the corresponding definition for generalized linear models. However, the familiar property that there exists a sufficient statistic having the same dimension as β no longer holds in general because curvature is introduced by the non-linearity of the predictor η_r. In the canonical-link case, $D_r = \lambda_r^{-1} \Sigma_r$, so $W_r^{-1} = \lambda_r^2 \Sigma_r^{-1}$ and

$$\left\{ \left(D_r \Sigma_r^{-1} \right)_s \otimes 1_q \right\} \mathcal{D}^2 \left(\mu_r ; \eta_r \right) = \lambda_r^{-1} \mathcal{D}^2 \left(\mu_{rs} ; \eta_r \right) = \lambda_r^{-3} K_{rs} \,,$$

where K_{rs} denotes the sth block of rows of K_r, $s \in \{1, \ldots, q\}$, with K_r the blocked $q^2 \times q$ matrix of third-order cumulants of the random vector Y_r. Hence expression (7) is considerably simplified, taking the form

$$U_t^* = U_t + \frac{1}{2} \sum_r \sum_{s=1}^{q} \lambda_r^{-2} \operatorname{tr} \left\{ H_r \Sigma_r^{-1} K_{rs} + \lambda_r F^{-1} (\Sigma_{rs} \otimes 1_p) \mathcal{D}^2 \left(\eta_r ; \beta \right) \right\} x_{rst}^* \,. \tag{8}$$

3.3 Special case: multivariate generalized linear model

Generalized linear models have the form (3), with $\eta_r(\beta) = Z_r \beta$, where the $q \times p$ design matrix Z_r is some appropriate function of a covariate vector and does not depend on β. Thus, in

equation (7), $\mathcal{D}^2(\eta_r;\beta) = 0$ and the adjusted score reduces to

$$U_t^* = U_t + \frac{1}{2}\sum_r \sum_{s=1}^{q} \mathrm{tr}\left[H_r W_r^{-1}\left\{\left(D_r\Sigma_r^{-1}\right)_s \otimes 1_q\right\}\mathcal{D}^2(\mu_r;\eta_r)\right] x_{rst}^* \,. \tag{9}$$

By the same argument, for canonical link functions expression (8) simplifies to

$$U_t^* = U_t + \frac{1}{2}\sum_r \sum_{s=1}^{q} \lambda_r^{-1}\,\mathrm{tr}\left(H_r\Sigma_r^{-1}K_{rs}\right) x_{rst}^* \,.$$

If the link is canonical the Fisher and observed information coincide, so that e_{ts} simplifies to the (t,s)th element of $1_p + RF^{-1}$, with 1_p the $p \times p$ identity matrix.

3.4 Notes on implementation

An obvious way of solving the adjusted score equation is a modified Fisher scoring iteration, in which the likelihood score is replaced by the adjusted score; the maximum likelihood estimates, if available and finite, often provide a good choice of starting values. However, more convenient and possibly more efficient schemes can be constructed by exploiting the special structure of the adjusted score for any given model. For example, Firth (1992a,b), in the special case of univariate generalized linear models with canonical link, derived a modified iterative re-weighted least squares procedure. In later sections this problem is revisited in order to generalize such procedures to univariate generalized linear and nonlinear models with any link function.

A generalized linear model with non-canonical link can be always expressed as a generalized nonlinear model with canonical link, and consequently (9) is equivalent to (8) for the corresponding generalized nonlinear model with canonical link. This fact can sometimes be exploited to allow the use of existing software, with relatively minor modifications, for implementation of the bias-reducing adjustments.

4 Univariate generalized linear models

4.1 Adjusted score

Suppose now that the response variable is scalar, i.e., $q = 1$. For notational simplicity, in the univariate case write $\kappa_{2,r}$ and $\kappa_{3,r}$ for the variance and the third cumulant of Y_r, respectively, and $w_r = d_r^2/\kappa_{2,r}$ for the working weights, where $d_r = \mathrm{d}\mu_r/\mathrm{d}\eta_r$ $(r = 1,\ldots,n)$. For a univariate generalized linear model with general link function, the adjusted score components (9) reduce to

$$U_t^* = U_t + \frac{1}{2}\sum_r h_r \frac{d_r'}{d_r}\sum_{s=1}^{p} e_{ts}x_{rs} \quad (t = 1,\ldots,p)\ , \tag{10}$$

where $d_r' = \mathrm{d}^2\mu_r/\mathrm{d}\eta_r^2$ and x_{rt} is the (r,t)th element of the $n \times p$ design matrix X. The quantity h_r is the rth diagonal element of the projection matrix $H = XF^{-1}X^TW$, where $W = \mathrm{diag}(w_1,\ldots,w_n)$, and $F = X^TWX$. Note that d_r'/d_r depends solely on the link function.

If the link is canonical, $d_r = \lambda_r^{-1}\kappa_{2,r}$, $d_r' = \lambda_r^{-2}\kappa_{3,r}$ and e_{ts} simplifies to the (t,s)th element of $1_p + RF^{-1}$ in the latter expression. Furthermore, if R is taken to be a matrix of zeros then, as given in Firth (1992a,b), $U_t^* = U_t + \sum_r h_r\{\kappa_{3,r}/(2\lambda_r\kappa_{2,r})\}x_{rt}$.

4.2 Existence of penalized likelihoods for univariate generalized linear models

For a generalized linear model with canonical link, the bias-reducing score adjustment corresponds to penalization of the likelihood by the Jeffreys (1946) invariant prior (Firth, 1993). More generally, in models with non-canonical link and $p \geq 2$, there need not exist a penalized log-likelihood l^* such that $\nabla l^*(\beta) \equiv U^*(\beta)$. The following theorem identifies those non-canonical link functions which always — i.e, regardless of the dimension or structure of X — do admit such a penalized likelihood interpretation.

Theorem 4.1 *In the class of univariate generalized linear models, there exists a penalized log-likelihood l^* such that $\nabla l^*(\beta) \equiv U(\beta) + A^{(E)}(\beta)$, for all possible specifications of design matrix X, if and only if the inverse link derivatives $d_r = 1/g_r'(\mu_r)$ satisfy*

$$d_r \equiv \alpha_r \kappa_{2,r}^{\omega} \quad (r = 1, \ldots, n)\,, \tag{11}$$

where α_r $(r = 1, \ldots, n)$ and ω do not depend on the model parameters. When condition (11) holds, the penalized log-likelihood is

$$l^*(\beta) = \begin{cases} l(\beta) \;+\; \dfrac{1}{4}\displaystyle\sum_r \log \kappa_{2,r}(\beta)^{h_r} & (\omega = 1/2) \\ \\ l(\beta) \;+\; \dfrac{\omega}{4\omega - 2} \log |F(\beta)| & (\omega \neq 1/2)\,. \end{cases} \tag{12}$$

The proof is in the Appendix.

In the above theorem the canonical link is the special case $\omega = 1$. With $\omega = 0$, condition (11) refers to identity links for which the log-likelihood penalty is identically zero. The case $\omega = 1/2$ is special on account of the fact that the working weights, and hence F and H, do not in that case depend on β.

Example 4.1 *Consider a Poisson generalized linear model with link function from the power family $\eta = (\mu^\nu - 1)/\nu$ (McCullagh & Nelder, 1989, § 2.2.3). Then $d_r = \mu_r^{1-\nu}$ and $\kappa_{2,r} = \mu_r$ $(r = 1, \ldots, n)$. Bias reduction based on expected information is equivalent to maximization of penalized likelihood (12) with $\omega = 1 - \nu$.*

Theorem 4.1 has direct practical consequences, notably for the construction of confidence sets. The use of profile penalized likelihood as suggested for example in Heinze & Schemper (2002) and Bull et al. (2007) is always available for a generalized linear model whose link function satisfies condition (11), but typically is not possible otherwise. Models which fail to meet condition (11) include, for example, probit and complementary log-log models for binary responses.

4.3 Pseudo-responses and implementation via modified working observations

The likelihood score components for univariate generalized linear models with general link have the form

$$U_t = \sum_r \frac{w_r}{d_r}(y_r - \mu_r)x_{rt} \quad (t = 1, \ldots, p)\;.$$

A simple substitution in expression (10) reveals an important feature of the adjusted score, which was recognised previously (Firth, 1993) in the more specific context of canonical-link

models. Consider the case of adjustments based on the expected information, i.e., e_{ts} is unity if $t = s$ and zero otherwise. If $h_r d'_r/(2w_r)$ $(r = 1, \ldots, n)$ were known constants (for example, when $\nu = 1/2$ in Example 4.1) then the bias reduction method would be formally equivalent to maximum likelihood with the pseudo-responses

$$y_r^* = y_r + \frac{1}{2} h_r \frac{d'_r}{w_r} \quad (r = 1, \ldots, n)$$

used in place of y_r. Table 1 gives the form of the pseudo-responses for some commonly used generalized linear models. The pseudo-responses suggest a simple approach to implementation, using a familiar likelihood-maximization algorithm such as iterative re-weighted least squares but with y_r replaced by y_r^*. Note that in general $h_r d'_r/w_r$ depends on the parameters, and so the value of y_r^* will be updated according to the current estimates at each step of the algorithm. This modified iterative re-weighted least squares step is equivalent to the modified Fisher scoring algorithm of Section 3.4 and can be more conveniently described in terms of the modified working observations

$$\zeta_r^* = \eta_r + \frac{y_r^* - \mu_r}{d_r} = \zeta_r - \xi_r \quad (r = 1, \ldots, n) \; .$$

Here $\zeta_r = \eta_r + (y_r - \mu_r)/d_r$ is the working observation for maximum likelihood, and $\xi_r = -d'_r h_r/(2w_r d_r)$, as defined in Cordeiro & McCullagh (1991). Thus, if the working observation ζ_r is modified by adding $-\xi_r$, the resulting iterative re-weighted least squares scheme returns the bias-reduced estimates.

This result is a direct consequence of the initial definition of the modifications in terms of the $O(n^{-1})$ bias of the maximum likelihood estimator and the fact that, as shown in Cordeiro & McCullagh (1991), the vector of $O(n^{-1})$ biases can be written as

$$n^{-1} b_1 = (X^T W X)^{-1} X^T W \xi \, .$$

These appealingly simple results do not extend in any obvious way to the case of multivariate responses, on account of the fact that W is no longer diagonal but rather is block diagonal, and because the vector of $O(n^{-1})$ biases can be expressed at best as a function of traces of products of matrices with no other apparent simplification.

5 Univariate generalized nonlinear models

5.1 Adjusted score

With $q = 1$ in expression (7), and under the notational conventions of the previous section, the adjusted score components for a univariate exponential family nonlinear model with known dispersion are

$$U_t^* = U_t + \frac{1}{2} \sum_r \left[h_r \frac{d'_r}{d_r} + w_r \operatorname{tr} \left\{ F^{-1} \mathcal{D}^2 \left(\eta_r ; \beta \right) \right\} \right] x_{rt}^* \quad (t = 1, \ldots, p) \; . \tag{13}$$

In the above expression, $x_{rt}^* = \sum_{s=1}^p e_{ts} x_{rs}$, $x_{rs} = \partial \eta_r / \partial \beta_s$ and $\mathcal{D}^2 \left(\eta_r ; \beta \right)$ is the $p \times p$ Hessian matrix of η_r with respect to β. The remaining quantities in (13) are as for generalized linear models.

For canonical links, in expression (13), $d'_r/d_r = \lambda^{-1} \kappa_{3,r}/\kappa_{2,r}$ and $w_r = \lambda_r^{-2} \kappa_{2,r}$.

Table 1: Pseudo-responses for several commonly used GLMs.

Distribution of Y	Link function $\eta = g(\mu)$	Pseudo-responses $y^* = y + hd'/(2w)$
Binomial (m, π)	$\eta = \log\{\pi/(1-\pi)\}$ †	$y^* = y + h(1/2 - \pi)$
	$\eta = \Phi^{-1}(\pi)$	$y^* = y - h\pi(1-\pi)\eta/\{2\phi(\eta)\}$
	$\eta = \log\{-\log(1-\pi)\}$	$y^* = y + h\pi(1-e^{\eta})/(2e^{\eta})$
	$\eta = -\log\{-\log(\pi)\}$	$y^* = y + h(1-\pi)(e^{-\eta}-1)/(2e^{-\eta})$
Poisson (μ)	$\eta = \log\mu$ †	$y^* = y + h/2$
	$\eta = \mu$	$y^* = y$
Gamma (μ, ν) $\mathrm{var}\,(Y) = \mu^2/\nu$	$\eta = -1/\mu$ †	$y^* = y + h\mu/\nu$
	$\eta = \log\mu$	$y^* = y + h\mu/(2\nu)$
	$\eta = \mu$	$y^* = y$
Inverse Gaussian (λ, μ)	$\eta = -1/(2\mu^2)$ †	$y^* = y + 3h\lambda\mu^2/2$

The normal distribution function and density are denoted by Φ and ϕ, respectively.

† Canonical link. The pseudo-responses for Binomial, Poisson and Gamma models with canonical link are also given in Firth (1992a,b).

5.2 Implementation

In the case of adjustments based on the expected information, the fitting procedures for univariate generalized linear models can be used in the nonlinear case with only slight changes to the definition of the pseudo-responses and the modified working observations. By (13), the pseudo-responses of Table 1 are straightforwardly adapted to the nonlinear case by adding the extra term $d_r \,\mathrm{tr}\left\{F^{-1}\mathcal{D}^2\left(\eta_r;\beta\right)\right\}/2$. The $O(n^{-1})$ bias of the maximum-likelihood estimator of the parameter vector in the case of a nonlinear predictor can be written in the form

$$n^{-1}b_1 = (X^T W X)^{-1} X^T W \xi^{(N)} ,$$

which generalizes the corresponding expression for normal nonlinear regression models given in Cook et al. (1986). The vector $\xi^{(N)}$ has components

$$\xi_r^{(N)} = -\frac{1}{2}\left[\frac{h_r d_r'}{w_r d_r} + \mathrm{tr}\left\{F^{-1}\mathcal{D}^2\left(\eta_r;\beta\right)\right\}\right] \quad (r = 1, \ldots, n) ,$$

and by similar arguments to those used for generalized linear models, the rth modified working variate takes the form $\zeta_r^* = \zeta_r - \xi_r^{(N)}$.

6 Illustration: Bias reduction for the RC(1) association model

6.1 The RC(1) association model

The utility of the preceding developments and the properties of the bias-reduced estimator are illustrated here by application to the RC(1) association model for the analysis of two-way contingency tables (Goodman, 1979, 1985).

Consider a two-way contingency table of cross-classified factors X and Y with R and S levels, respectively. The entries of the table are assumed to be realizations of independent

Poisson random variables Y_{rs} with means μ_{rs} $(r = 1, \ldots, R; s = 1 \ldots, S)$. For the RC(1) model, μ_{rs} is linked to a nonlinear function η_{rs} of model parameters according to

$$\log \mu_{rs} = \eta_{rs} = \lambda + \lambda_r^X + \lambda_s^Y + \rho\gamma_r\delta_s \,, \tag{14}$$

where λ_r^X and λ_s^Y are the row and column main effects, γ_r and δ_s are row and column score parameters, and ρ is an association parameter. The RC(1) model is viewed here as a univariate generalized nonlinear model with canonical link.

Write the parameter vector as

$$\beta = \left(\lambda, \lambda_1^X, \ldots, \lambda_R^X, \lambda_1^Y, \ldots, \lambda_S^Y, \rho, \gamma_1, \ldots, \gamma_R, \delta_1, \ldots, \delta_S\right)^T \,,$$

with $p = 2(R+S+1)$. The full parameter vector β is unidentified, a set of 6 constraints being required in order to fix the location of $\left\{\lambda_r^X\right\}$ and $\left\{\lambda_s^Y\right\}$ and the location and scale of $\{\gamma_r\}$ and $\{\delta_s\}$. In what follows, it will be assumed for simplicity that 6 suitably chosen elements of β are constrained at fixed values.

6.2 Modified working variates

Let $\eta_r = (\eta_{r1}, \ldots, \eta_{rS})^T$ and denote the reduced vector of parameters, after removal of the constrained elements, by β^-. For the RC(1) model, $d'_{rs}/d_{rs} = 1$ and $w_{rs} = \mu_{rs}$. Thus, by the results in Section 5.2, the vector of modified working variates $\zeta^* = (\zeta_{11}^*, \ldots, \zeta_{1S}^*, \ldots, \zeta_{R1}^*, \ldots, \zeta_{RS}^*)$ has components

$$\zeta_{rs}^* = \zeta_{rs} + \frac{h_{rs}}{2\mu_{rs}} + M_{rs} \,. \tag{15}$$

Here $M_{rs} = \mathrm{tr}\left\{F^{-1}\mathcal{D}^2\left(\eta_r; \beta^-\right)\right\}/2$, and h_{rs} is the sth diagonal element of the $S \times S$ matrix $X_r F^{-1} X_r^T W_r$, with W_r a diagonal matrix with sth diagonal element μ_{rs}, $s = 1, \ldots, S$ and with X_r the $S \times (p-6)$ matrix of derivatives of η_r with respect to β^-.

The sth row of X_r results from the deletion of all components of $\mathcal{D}\left(\eta_{rs}; \beta\right)$ which correspond to constrained parameters. The derivatives are

$$\mathcal{D}\left(\eta_{rs}; \beta\right) = \left(1, i_r^R, i_s^S, \gamma_r\delta_s, \rho\delta_s i_r^R, \rho\gamma_r i_r^R\right) ,$$

with, for example, i_r^R denoting a $1 \times R$ row vector of zeros with 1 at the rth position. Thus, by noting that $\mathcal{D}^2\left(\eta_r; \beta^-\right) = \mathcal{D}\left(X_r; \beta^-\right)$, and after some straightforward algebra, M_{rs} can be expressed conveniently in the form

$$M_{rs} = \gamma_r C(\rho, \delta_s) + \delta_s C(\rho, \gamma_r) + \rho C(\gamma_r, \delta_s) \,. \tag{16}$$

Here $C(\kappa, \nu)$ for any given pair of unconstrained parameters κ and ν denotes the corresponding element of F^{-1}; if either of κ or ν is constrained, $C(\kappa, \nu) = 0$.

Expressions (15) and (16) combine to provide a rather simple implementation of the bias-reduction method in this case via the modified iterative weighted least squares procedure.

6.3 Simulation study

For illustration, consider the mental health status data given in Agresti (2002, Table 9.9). The subjects are cross-classified according to two factors: parents' socioeconomic status (X), which has 6 levels $(R = 6)$ and mental health status (Y) which has four levels $(S = 4)$. The constraints used are

$$\lambda_1^X = \lambda_1^Y = 0 \,, \quad \gamma_1 = \delta_1 = -1 \quad \text{and} \quad \gamma_6 = \delta_4 = 1 \,.$$

Table 2: Results for the mental health status data.

	Estimates		Simulation results					
			Bias ($\times 10^2$)		MSE[3] ($\times 10$)		Coverage (%)	
	ML[1]	BR[2]	ML	BR	ML	BR	ML	BR
λ	3.773	3.784	−1.063	−0.022	0.109	0.106	94.7	94.8
λ_2^X	−0.067	−0.067	−0.022	0.012	0.084	0.083	95.3	95.3
λ_3^X	0.090	0.087	0.229	0.050	0.083	0.082	95.2	95.3
λ_4^X	0.374	0.370	0.345	0.051	0.082	0.081	95.0	95.1
λ_5^X	−0.033	−0.034	0.091	0.082	0.133	0.130	94.6	94.8
λ_6^X	−0.281	−0.278	−0.436	0.024	0.203	0.195	94.2	94.6
λ_2^Y	0.802	0.793	0.884	0.011	0.122	0.117	94.6	94.6
λ_3^Y	0.310	0.302	0.774	0.016	0.150	0.144	94.5	94.6
λ_4^Y	0.430	0.426	0.478	−0.034	0.247	0.238	94.3	94.5
ρ	0.377	0.374	0.241	−0.067	0.060	0.058	94.2	94.5
γ_2	−1.006	−0.974	−3.882	0.296	1.755	1.459	95.3	95.3
γ_3	−0.494	−0.482	−1.628	0.110	1.252	1.108	95.6	96.2
γ_4	−0.222	−0.220	−0.573	−0.015	1.006	0.906	95.4	96.2
γ_5	0.449	0.429	2.277	−0.297	1.497	1.299	95.1	95.6
δ_2	−0.005	0.001	−0.574	−0.119	0.627	0.586	93.1	94.4
δ_3	0.174	0.180	−0.250	−0.135	0.796	0.747	93.5	94.8

[1] ML: maximum likelihood, [2] BR: Bias reduction method, [3] MSE: Mean squared error

With these constraints, the maximum likelihood estimates are obtained, and 500000 tables are simulated from the maximum likelihood fit. For each simulated table the following are computed: the maximum likelihood estimates, the bias-reduced estimates, corresponding estimated standard errors computed as square roots of the diagonal of the inverse of the Fisher information, and a nominally 95% Wald-type confidence interval for each parameter separately. The resulting estimated bias, mean squared error and interval coverage probability for each parameter are given in Table 2; simulation error is small enough in every case to allow high confidence in the last digit of the reported estimates.

In Table 2, in every case the estimated absolute bias and mean squared error are smaller for the bias-reduced estimator than for maximum likelihood. Due to the large sample size in this example the Wald-type confidence intervals from both estimates have coverage rates close to the nominal level; the intervals based on bias-reduced estimates are slightly more conservative.

It should be noted that, strictly speaking, the estimated biases, variances and mean squared errors reported here for the maximum likelihood estimator are all conditional upon the finiteness of the maximum likelihood estimates. Since the maximum likelihood estimator for each parameter in the RC(1) model has non-zero probability of being infinite, the unconditional moments are all undefined. In the case of the mental health status data, the large sample size makes the probability of an infinite-valued maximum likelihood estimate negligible; indeed, not a single instance of monotone likelihood was seen among the 500000 tables studied above. With small or moderate-sized samples, however, the possibility of infinite estimates cannot be neglected. To examine this, a parallel simulation study, the detailed results of which will not be reported here, was also made of a table with much smaller counts.

For the cross-classification of 135 women by periodontal condition and calcium intake given in Goodman (1981, Table 1a), in about 4% of samples at least one of the the maximum likelihood estimates for the RC(1) model was found to be infinite-valued. In contrast, the bias-reduced estimates were always finite.

7 Concluding remarks

Explicit, general formulae have been derived for the adjusted score equations that produce second-order unbiased estimators, starting from the wide class of multivariate-response exponential family nonlinear models and narrowing down to the simplest case of canonically-linked generalized linear models. As was shown in the case of the RC(1) model, simplification of the formulae is also possible in some other special cases, such as generalized bilinear models, by exploiting the specific structure of the various quantities involved.

The apparent focus here on models with known dispersion does not affect the wide applicability of the results. In generalized linear and nonlinear models where the dispersion λ is unknown, it is usually estimated separately from the parameters β which determine the mean; given an estimate of λ, the methods developed here can simply be applied with λ fixed at its current estimate. The well known orthogonality of mean and dispersion parameters plays an important role in this.

Generally, as for the maximum likelihood estimator, approximate confidence intervals for the bias-reduced estimator can be constructed by the usual Wald method. However, as noted in Heinze & Schemper (2002) and Bull et al. (2007), for logistic regressions the Wald-type intervals can have poor coverage properties. Heinze & Schemper (2002) and Bull et al. (2007) propose the use, instead, of intervals based on the profile penalized likelihood, which are found to have better properties. Theorem 4.1 shows that such use of profile penalized likelihood can be extended beyond logistic regressions but is not universally available.

Acknowledgement

The authors thank the Editor, the Associate Editor and a referee for helpful comments which have substantially improved the presentation. Financial support from the Engineering and Physical Sciences Research Council (IK) and the Economic and Social Research Council (DF) is gratefully acknowledged.

Appendix

Proof of Theorem 4.1. Note that $d_r'/d_r = \mathrm{dlog}\, d_r/\mathrm{d}\eta_r$ and so, for adjustments based on the expected information, (10) can be written as

$$U_t^* = U_t + \frac{1}{2}\,\mathrm{tr}\,(HET_t) \quad (t = 1, \ldots, p)\ , \tag{17}$$

with $E = \mathrm{diag}(\mathrm{dlog}\, d_1/\mathrm{d}\eta_1, \ldots, \mathrm{dlog}\, d_n/\mathrm{d}\eta_n)$ and $T_t = \mathrm{diag}(x_{1t}, \ldots, x_{nt})$. As, for example, in the case of the existence of quasi-likelihoods in McCullagh & Nelder (1989, § 9.3.2), there exists a penalized log-likelihood l^* such that $\nabla l^*(\beta) \equiv U(\beta) + A^{(E)}(\beta)$, if and only if $\partial U_s^*(\beta)/\partial\beta_t = \partial U_t^*(\beta)/\partial\beta_s$ for every $s, t \in \{1, 2, \ldots, p\}$, $s \neq t$. By (17), this holds if and only if

$$\frac{\partial \mathrm{tr}\,(HET_t)}{\partial\beta_s} = \mathrm{tr}\left(H\frac{\partial E}{\partial\beta_s}T_t\right) + \mathrm{tr}\left(\frac{\partial H}{\partial\beta_s}ET_t\right)$$

is invariant under interchange of the subscripts s and t. The first term in the right hand side of the above expression is invariant because $\partial E/\partial\beta_s = ET_s$ and T_s and T_t are by definition diagonal so that matrix multiplication is commutative for them. For the second term,

$$\frac{\partial H}{\partial \beta_s} = -X(X^TWX)^{-1}X^TW_sX(X^TWX)^{-1}X^TW + X(X^TWX)^{-1}X^TW_s\,,$$

where $W_s = \partial W/\partial\beta_s = W(2E-\Lambda)T_s$ with $\Lambda = \mathrm{diag}(\mathrm{dlog}\,\kappa_{2,1}/\mathrm{d}\eta_1, \ldots, \mathrm{dlog}\,\kappa_{2,n}/\mathrm{d}\eta_n)$. Thus,

$$\mathrm{tr}\left(\frac{\partial H}{\partial \beta_s}ET_t\right) = 2\,\mathrm{tr}\,(HET_sET_t) - 2\,\mathrm{tr}\,(HET_sHET_t) - \mathrm{tr}\,(H\Lambda T_sET_t) + \mathrm{tr}\,(H\Lambda T_sHET_t)\ .$$

By the properties of the trace function, the first three terms in the right hand side of the above expression are invariant under interchange of s and t. Thus the condition is reduced to the invariance of $\mathrm{tr}(H\Lambda T_sHET_t)$. The projection matrix H can be written as $H = SW$, with $S = X(X^TWX)^{-1}X^T$. Let $\tilde{H} = W^{1/2}SW^{1/2}$. Taking into account the symmetry of $\tilde{H}$, some trivial algebra and an application of Theorem 3 of Magnus & Neudecker (1999, Chapter 2) gives

$$\mathrm{tr}\,(H\Lambda T_sHET_t) = \mathrm{tr}\left(\tilde{H}T_s\Lambda\tilde{H}ET_t\right) = (\mathrm{vec}\,T_t)^T\left\{\left(E\tilde{H}\Lambda\right)\otimes\tilde{H}\right\}\mathrm{vec}\,T_s\,. \qquad (18)$$

The columns of X are by assumption linearly independent and thus (18) is invariant under interchanges of s and t if and only if $a^T\{(E\tilde{H}\Lambda)\otimes\tilde{H}\}b$ is a symmetric bilinear form for distinct vectors a and b of appropriate dimension, or equivalently if and only if $E\tilde{H}\Lambda$ is symmetric, i.e.,

$$\frac{\mathrm{dlog}\,d_r}{\mathrm{d}\eta_r}\frac{\mathrm{dlog}\,\kappa_{2,i}}{\mathrm{d}\eta_i}\tilde{h}_{ri} = \frac{\mathrm{dlog}\,\kappa_{2,r}}{\mathrm{d}\eta_r}\frac{\mathrm{dlog}\,d_i}{\mathrm{d}\eta_i}\tilde{h}_{ri} \quad (r,i=1,\ldots,n; r>i)\,, \qquad (19)$$

with $\tilde{h}_{ri}$ the (r,i)th element of $\tilde{H}$.

In general the above equations are not satisfied simultaneously, except possibly for special structures of the design matrix X, which cause $\tilde{h}_{ri} = 0$ for a set of pairs (r,i). Hence, assuming that $\tilde{h}_{ri} \neq 0$ $(r,i=1,\ldots,n; r>i)$, the general equation in (19) reduces to $\mathrm{dlog}\,d_r/\mathrm{d}\eta_r\mathrm{dlog}\,\kappa_{2,i}/\mathrm{d}\eta_i = \mathrm{dlog}\,\kappa_{2,r}/\mathrm{d}\eta_r\mathrm{dlog}\,d_i/\mathrm{d}\eta_i$. Now, if $\mathrm{dlog}\,\kappa_{2,r}/\mathrm{d}\eta_r = \mathrm{dlog}\,\kappa_{2,i}/\mathrm{d}\eta_i = 0$ for some pair (r,i) then the equation for this (r,i) is satisfied. Thus, without loss of generality assume that $\mathrm{dlog}\,\kappa_{2,r}/\mathrm{d}\eta_r \neq 0$ for every $r \in \{1,\ldots,n\}$. Under these assumptions condition (19) can be written as $\mathrm{dlog}\,d_r/\mathrm{d}\eta_r = \omega\mathrm{dlog}\,\kappa_{2,r}/\mathrm{d}\eta_r$ $(r=1,\ldots,n)$, where ω does not depend on the model parameters. By integration of both sides of $\mathrm{dlog}\,d_r/\mathrm{d}\eta_r = \omega\mathrm{dlog}\,\kappa_{2,r}/\mathrm{d}\eta_r$ with respect to η_r, a necessary condition for the adjusted score to be the gradient of a penalized likelihood is thus

$$d_r \equiv \alpha_r\kappa_{2,r}^{\omega} \quad (r=1,\ldots,n)\ , \qquad (20)$$

where $\{\alpha_r : r=1,\ldots,n\}$ are real constants not depending on the model parameters.

To check that (20) is sufficient, simply note that if we substitute accordingly, the matrix $E\tilde{H}\Lambda$ is symmetric.

In addition, if condition (20) is satisfied for some ω and α_r $(r=1,\ldots,n)$ then the rth diagonal element of E is $\mathrm{dlog}\,d_r/\mathrm{d}\eta_r = \omega\kappa'_{2,r}/\kappa_{2,r}$ for every $r \in \{1,\ldots,n\}$, with $\kappa'_{2,r} = \mathrm{d}\kappa_{2,r}/\mathrm{d}\eta_r$. On the other hand, $\mathrm{d}w_r/\mathrm{d}\eta_r = (2\omega-1)w_r\kappa'_{2,r}/\kappa_{2,r}$. Hence, by (17) and for $\omega \neq 1/2$ the tth component of the adjusted score vector is

$$U_t(\beta) + \frac{\omega}{4\omega-2}\,\mathrm{tr}\left\{X\left(X^TW(\beta)X\right)^{-1}X^TW_t(\beta)\right\} = \frac{\partial}{\partial\beta_t}\left\{l(\beta) + \frac{\omega}{4\omega-2}\log\left|X^TW(\beta)X\right|\right\}\,,$$

where and $W_t(\beta) = \partial W(\beta)/\partial\beta_t$.

If $\omega = 1/2$ then $w_r = \alpha_r^2$ $(r = 1, \ldots, n)$. Hence, the hat matrix H does not depend on the model parameters. Thus, by (10), the tth component of the adjusted score vector is

$$U_t(\beta) + \frac{1}{4}\sum_r h_r \frac{\kappa'_{2,r}(\beta)}{\kappa_{2,r}(\beta)} x_{rt} = \frac{\partial}{\partial \beta_t}\left\{ l(\beta) + \frac{1}{4}\sum_r \log \kappa_{2,r}(\beta)^{h_r} \right\}.$$

References

Agresti, A. (2002). *Categorical Data Analysis*. New York: Wiley.

Bull, S. B., Lewinger, J. B. & Lee, S. S. F. (2007). Confidence intervals for multinomial logistic regression in sparse data. *Statistics in Medicine* **26**, 903–918.

Bull, S. B., Mak, C. & Greenwood, C. (2002). A modified score function estimator for multinomial logistic regression in small samples. *Computational Statistics and Data Analysis* **39**, 57–74.

Cook, R. D., Tsai, C.-L. & Wei, B. C. (1986). Bias in nonlinear regression. *Biometrika* **73**, 615–623.

Cordeiro, G. M. & McCullagh, P. (1991). Bias correction in generalized linear models. *Journal of the Royal Statistical Society, Series B: Methodological* **53**, 629–643.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton: Princeton University Press.

Fahrmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York: Springer.

Firth, D. (1992a). Bias reduction, the Jeffreys prior and GLIM. In *Advances in GLIM and Statistical Modelling: Proceedings of the GLIM 92 Conference, Munich*, L. Fahrmeir, B. Francis, R. Gilchrist & G. Tutz, eds. New York: Springer.

Firth, D. (1992b). Generalized linear models and Jeffreys priors: An iterative generalized least-squares approach. In *Computational Statistics I*, Y. Dodge & J. Whittaker, eds. Heidelberg: Physica-Verlag.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38.

Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association* **74**, 537–552.

Goodman, L. A. (1981). Association models and canonical correlation in the analysis of cross-classifications having ordered categories. *Journal of the American Statistical Association* **76**, 320–334.

Goodman, L. A. (1985). The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *The Annals of Statistics* **13**, 10–69.

Heinze, G. & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine* **21**, 2409–2419.

Heinze, G. & Schemper, M. (2004). A solution to the problem of monotone likelihood in Cox regression. *Biometrics* **57**, 114–119.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London* **186**, 453–461.

Magnus, J. R. & Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Chichester: Wiley.

McCullagh, P. & Nelder, J. (1989). *Generalized Linear Models*. London: Chapman and Hall, 2nd ed.

Mehrabi, Y. & Matthews, J. N. S. (1995). Likelihood-based methods for bias reduction in limiting dilution assays. *Biometrics* **51**, 1543–1549.

Pettitt, A. N., Kelly, J. M. & Gao, J. T. (1998). Bias correction for censored data with exponential lifetimes. *Statistica Sinica* **8**, 941–964.

Sartori, N. (2006). Bias prevention of maximum likelihood estimates for scalar skew normal and skew t distributions. *Journal of Statistical Planning and Inference* **136**, 4259–4275.

Wei, B. (1997). *Exponential Family Nonlinear Models*. New York: Springer-Verlag Inc.

Zorn, C. (2005). A solution to separation in binary response models. *Political Analysis* **13**, 157–170.