Ann. Inst. Statist. Math.
Vol. 54, No. 2, 355–366 (2002)

ON THE POSITIVE DEFINITENESS OF THE INFORMATION MATRIX UNDER THE BINARY AND POISSON MIXED MODELS*

RAHUL MUKERJEE[1] AND BRAJENDRA C. SUTRADHAR[2]

[1] *Indian Institute of Management, Post Box No. 16757, Calcutta 700 027, India,* e-mail: rmuk1@hotmail.com
[2] *Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, Newfoundland, Canada A1C 5S7,* e-mail: bsutradh@math.mun.ca

(Received January 11, 2000; revised January 23, 2001)

Abstract. Binary and Poisson generalized linear mixed models are used to analyse over/under-dispersed proportion and count data, respectively. As the positive definiteness of the information matrix is a required property for valid inference about the fixed regression vector and the variance components of the random effects, this paper derives the relevant necessary and sufficient conditions under both these models. It is found that the conditions for the positive definiteness are not identical for these two nonlinear mixed models and that a mere analogy with the usual linear mixed model does not dictate these conditions.

Key words and phrases: Estimating function, Fisher information matrix, generalised linear mixed model, joint estimates, likelihood estimation, positive definiteness, regression effects, variance component of the random effects.

1. Introduction

Generalised linear models (Nelder and Wedderburn (1972), McCullagh and Nelder (1989)) are used to analyse a wide variety of discrete and continuous responses that can be assumed to be independent. In many problems, however, responses are clustered and they are likely to be correlated. This within-cluster correlation must be taken into account to correctly assess the relation between the responses and the possible covariates. For example, in a genetic epidemiology study, observations on members of the same family are usually correlated and the correlations may be modeled by introducing the same random cluster effect over the individuals of the cluster.

In this paper, we deal with mixed effects models which will reflect heterogeneity across clusters in the regression coefficients, causing observations from the same cluster to be correlated. Note that the mixed effects generalized linear model is composed of a response y_{ij} and a vector of p predictors x_{ij} for observations $j = 1, \dots, n_i$ within clusters $i = 1, \dots, K$. The within-cluster correlation arises from heterogeneity among clusters in the coefficients for a q-dimensional (say) additional set of covariates z_i. If, conditional on a random vector γ_i, y_{ij} follows an exponential family distribution of the form

$$f(y_{ij} \mid \gamma_i) = \exp[\{y_{ij}\theta^*_{ij} - a(\theta^*_{ij})\}/\phi + b(y_{ij}, \phi)], \tag{1.1}$$

*This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada and Centre for Management and Development Studies, Indian Institute of Management, Calcutta.

where $\theta_{ij}^* = h(\eta_{ij})$ with $\eta_{ij} = x_{ij}^T\beta + z_i^T\gamma_i$, $a(\cdot)$, $b(\cdot)$ and $h(\cdot)$ are of known functional form, ϕ is a possibly unknown scale parameter, then y_{ij} unconditionally follows an exponential family based mixed model which is obtained by integrating over the distribution of γ_i. Under normality assumption of γ_i, i.e., $\gamma_i \sim^{\text{iid}} N_q(0, \Sigma)$, many authors, recently, have dealt with the above exponential family based mixed model mainly for the estimation of the regression vector β and the $q(q+1)/2$ distinct elements of the covariance matrix of the random effects, Σ. For example, we refer to Zeger and Karim (1991), Breslow and Clayton (1993), Waclawiw and Liang (1993), Breslow and Lin (1995), Lee and Nelder (1996), and Jiang (1998), among others.

Note that one of the main reasons for studying the exponential family based mixed model is that, apart from the normal mixed model, this model accommodates two highly practical models, namely, the binary and Poisson generalized linear mixed models (GLMMs). These latter two models, with some modifications in some cases, appear to fit many biomedical data, for example. Consequently, these binary and Poisson GLMMs have been studied by many authors over the last two decades. The binary GLMMs, for example, are studied by Williams (1982), Stiratelli *et al.* (1984), Moore (1986), and Karim and Zeger (1992). Similarly, for studies on Poisson GLMMs, we, for example, refer to Breslow (1984), Moore (1986), Lawless (1987), Morton (1987), and Sutradhar and Qu (1998).

Although many of the above mentioned authors have customarily considered the information matrix under the binary and/or Poisson GLMMs, no study on the conditions ensuring the positive definiteness (p.d.-ness) of the information matrix is as yet available. Note that, especially in small samples, just as in ordinary linear models, there is a risk that the information matrix may not be positive definite. This makes the issue of such p.d.-ness important since it is a required property for further valid inference about the fixed regression vector and the variance components of the random effects. While the absence of this p.d.-ness is sometimes revealed numerically through lack of convergence or through poor performance of the maximum likelihood estimators, we consider it appropriate to study the subject theoretically with a view to obtaining systematic results via specification of the relevant necessary and sufficient conditions. Verification of these necessary and sufficient conditions before one starts analyzing the data can entail considerable saving of efforts in the event of lack of p.d.-ness of the information matrix. For example, if these conditions do not hold then more data should be collected till such p.d.-ness is achieved even before any further analysis is attempted.

The necessary and sufficient conditions for the binary and Poisson GLMMs are provided in Sections 2 and 3 respectively. As the information matrix is not analytically tractable under either of these models, a subtle technique based on carefully chosen unbiased estimating functions is employed for this purpose. We conclude the paper in Section 4 where it is noted that, somewhat counterintuitively, mere analogy with the usual linear mixed model does not dictate the present results.

2. Binary generalized linear mixed models

Let y_{ij} be a binary response variable with associated fixed covariate vector $x_{ij} = (x_{ij1}, \ldots, x_{ijp})^T$. Then conditional on a random vector γ_i, the binary regression model assumes that $\Pr(y_{ij} = 1 \mid \gamma_i) = \{1 + \exp(-x_{ij}^T\beta - z_i^T\gamma_i)\}^{-1} = p_{ij}^*$, say. This conditional distribution is of the exponential form (1.1) with $\theta_{ij}^* = x_{ij}^T\beta + z_i^T\gamma_i$ and $a'(\theta_{ij}^*) = \{1 + \exp(-x_{ij}^T\beta - z_i^T\gamma_i)\}^{-1} = p_{ij}^*$. Similarly to Zeger and Karim (1991), Breslow and Lin

(1995), assume that $\gamma_i \sim^{\text{iid}} N_q(0, \Sigma)$, where $\Sigma = (\sigma_{uv})$ is an unknown positive definite (p.d.) dispersion matrix.

Let $n = \sum_{i=1}^{K} n_i$, $y = (y_{11}, \ldots, y_{1n_1}, \ldots, y_{K1}, \ldots, y_{Kn_K})^T$, $\theta = (\beta^T, \sigma^T)^T$, where the $\binom{q+1}{2} \times 1$ vector σ is defined as $\sigma = (\sigma_{11}, \ldots, \sigma_{1q}, \sigma_{22}, \ldots, \sigma_{2q}, \ldots, \sigma_{qq})^T$. Here y is the observation vector while θ is the vector of unknown parameters. Let $I \equiv I(\theta)$ be the Fisher information matrix for θ on the basis of y. We shall obtain necessary and sufficient conditions for $I(\theta)$ to be p.d. for every θ. Note that $I(\theta)$ is the very basis of any kind of inference about θ and, as in ordinary linear-models, lack of positive definiteness of $I(\theta)$ results in non-identifiability of the parameters.

Some more notation will help. Define the $n \times p$ matrix

$$X = [x_{11}, \ldots, x_{1n_1}, \ldots, x_{K1}, \ldots, x_{Kn_K}]^T.$$

For each i, let $z_i = (z_{i1}, \ldots, z_{iq})^T$ and define the $\binom{q+1}{2} \times 1$ vector

$$z_i^* = [z_{i1}^2, 2z_{i1}z_{i2}, \ldots, 2z_{i1}z_{iq}, z_{i2}^2, \ldots, 2z_{i2}z_{iq}, \ldots, z_{iq}^2]^T.$$

Also define the $K \times \binom{q+1}{2}$ matrix $Z = [z_1^*, \ldots, z_i^*, \ldots, z_K^*]^T$. Let $W = \{i : n_i \geq 2\}$ and w be the cardinality of W. If $w > 0$ then define $\tilde{Z}$ as the $w \times \binom{q+1}{2}$ submatrix of Z consisting of rows $(z_i^*)^T$ for $i \in W$. We are now in a position to present our main result for the binary GLMM, as in the following theorem.

THEOREM 2.1. *Under the binary generalized linear mixed model, the information matrix $I(\theta)$ is p.d. for all θ if and only if $w > 0$ and both X and $\tilde{Z}$ have full column rank.*

We first prove the 'only if' part of the theorem. For $1 \leq j \leq n_i$, $1 \leq i \leq K$, let $\mu_{ij} = x_{ij}^T\beta$, and for $1 \leq i \leq K$, let

$$\lambda_i \equiv \lambda_i(\sigma) = z_i^T \Sigma z_i = \sum_{u=1}^{q}\sum_{v=1}^{q} \sigma_{uv} z_{iu} z_{iv}.$$

Define the $n \times 1$ vector $\mu \equiv \mu(\beta) = (\mu_{11}, \ldots, \mu_{1n_1}, \ldots, \mu_{K1}, \ldots, \mu_{Kn_K})^T$, and the $K \times 1$ vector $\lambda \equiv \lambda(\sigma) = (\lambda_1, \ldots, \lambda_K)^T$. Also, let $\psi \equiv \psi(\theta) = (\mu^T, \lambda^T)^T$. It then follows that

$$\frac{\partial \mu^T}{\partial \beta} = X^T, \quad \frac{\partial \mu^T}{\partial \sigma} = 0, \quad \frac{\partial \lambda^T}{\partial \beta} = 0, \quad \frac{\partial \lambda^T}{\partial \sigma} = Z^T, \tag{2.1}$$

implying that

$$\frac{\partial \psi^T}{\partial \theta} = \begin{bmatrix} X^T & 0 \\ 0 & Z^T \end{bmatrix}. \tag{2.2}$$

We now consider the likelihood function of θ on the basis of y. Let $\phi_q(\cdot; \Sigma)$ denote the q-variate normal density with mean vector 0 and dispersion matrix Σ. As γ_i has the density $\phi_q(\cdot; \Sigma)$, then $\xi_i = z_i^T \gamma_i$ is univariate normal with mean 0 and variance $\lambda_i = z_i^T \Sigma z_i$. Consequently, the likelihood function for the binary mixed model is given by

$$L = \prod_{i=1}^{K} \int_{-\infty}^{\infty} \left[\prod_{j=1}^{n_i} \{\exp\{(\mu_{ij} + \xi_i) y_{ij}\} / [1 + \exp(\mu_{ij} + \xi_i)]\} \right] \phi_1(\xi_i; \lambda_i) d\xi_i. \tag{2.3}$$

It then follows that the likelihood function L depends on θ only through $\psi = (\mu^T, \lambda^T)^T$. Therefore, using (2.2),

$$\frac{\partial \log L}{\partial \theta} = \frac{\partial \psi^T}{\partial \theta}\frac{\partial \log L}{\partial \psi} = \begin{bmatrix} X^T & 0 \\ 0 & Z^T \end{bmatrix}\left(\frac{\partial \log L}{\partial \psi}\right). \tag{2.4}$$

Now for each $i (i = 1, \ldots, K)$,

$$\frac{\partial \log L}{\partial \lambda_i} = \frac{1}{2\lambda_i^2 L_i}\int_{-\infty}^{\infty}\left[\prod_{j=1}^{n_i}\{\exp\{(\mu_{ij}+\xi_i)y_{ij}\}/[1+\exp(\mu_{ij}+\xi_i)]\}\right] \times(\xi_i^2-\lambda_i)\phi_1(\xi_i;\lambda_i)d\xi_i, \tag{2.5}$$

where

$$L_i = \int_{-\infty}^{\infty}\left[\prod_{j=1}^{n_i}\{\exp\{(\mu_{ij}+\xi_i)y_{ij}\}/[1+\exp(\mu_{ij}+\xi_i)]\}\right]\phi_1(\xi_i;\lambda_i)d\xi_i.$$

Now by (2.5), for $i \notin W$, noting that y_{i1} is either 0 or 1 for the binary mixed model, we have

$$\left.\frac{\partial \log L}{\partial \lambda_i}\right|_{\beta=0} = \frac{1}{2\lambda_i^2 L_i}\int_{-\infty}^{\infty}\{\exp(\xi_i y_{i1})/[1+\exp(\xi_i)]\}(\xi_i^2-\lambda_i)\phi_1(\xi_i;\lambda_i)d\xi_i = 0. \tag{2.6}$$

In order to see why (2.6) should be true, let J_0 and J_1 be the integral in (2.6) corresponding to $y_{i1} = 0$ and $y_{i1} = 1$ respectively. Clearly $J_0 + J_1 = 0$. On the other hand, transforming $\xi_i = -\eta_i$ (say) in J_0, we get $J_0 = J_1$. Thus $J_0 = J_1 = 0$ which establishes (2.6).

If $w = 0$ then by (2.6), $\{\partial \log L/\partial\lambda\}_{\beta=0} = 0$, so that by (2.4),

$$\left.\frac{\partial \log L}{\partial \theta}\right|_{\beta=0} = \begin{bmatrix} X^T\left(\dfrac{\partial \log L}{\partial \mu}\right)_{\beta=0} \\ 0 \end{bmatrix}, \tag{2.7}$$

where the null vector appearing in the bottom of the right-hand side of (2.7) is of order $\binom{q+1}{2} \times 1$. On other hand, if $w > 0$ then define $\tilde{\lambda}$ as a $w \times 1$ vector with elements λ_i, $i \in W$, arranged in an order that corresponds to the ordering of the rows of $\tilde{Z}$. Then writing $\tilde{\psi} = (\mu^T, \tilde{\lambda}^T)^T$, by (2.4) and (2.6),

$$\left.\frac{\partial \log L}{\partial \theta}\right|_{\beta=0} = \begin{bmatrix} X^T & 0 \\ 0 & \tilde{Z}^T \end{bmatrix}\left(\frac{\partial \log L}{\partial \tilde{\psi}}\right)_{\beta=0}. \tag{2.8}$$

The "only if" part of Theorem 2.1 is immediate from (2.7) and (2.8).

We now proceed to prove the "if" part of the theorem. This is, however, more difficult since an analytical expression for $I(\theta)$ is intractable because of the integrals involved with regard to the random effects. As shown below, we have to start from a carefully chosen unbiased estimating function and follow an approach akin to that employed in the derivation of the Rao-Cramer bound for the multiparameter case.

Suppose that the conditions stated in Theorem 2.1 hold. Furthermore, without any loss of generality, let $W = \{1, \ldots, w\}$, where $1 \leq w \leq K$. Let

$$U = (y_{11}y_{12}, \ldots, y_{w1}y_{w2})^T, \tag{2.9}$$

and

$$g(y) = (y^T, U^T)^T. \tag{2.10}$$

In the conditional setup with $\gamma_1, \ldots, \gamma_K$ conditionally fixed, it is easily seen that the dispersion matrix of $g(y)$ is p.d. whatever β and $\gamma_1, \ldots, \gamma_K$ might be. As such, denoting the dispersion matrix of $g(y)$ in the unconditional set-up by $V(\theta)$, the following lemma is evident.

LEMMA 2.1. *The dispersion matrix $V(\theta)$ of $g(y)$ is p.d. for every θ.*

The following lemma is the most crucial one in our proof.

LEMMA 2.2. *Let $h(\theta) = E_\theta\{g(y)\}$. Under the conditions stated in Theorem 2.1, the matrix $\partial h(\theta)/\partial \theta^T$ has full column rank for every θ.*

The proof of the lemma is given in Appendix A.

LEMMA 2.3. *The matrix $I(\theta) - (\frac{\partial h(\theta)}{\partial \theta^T})^T \{V(\theta)\}^{-1} (\frac{\partial h(\theta)}{\partial \theta^T})$ is nonnegative definite.*

PROOF. Since the dispersion matrix of $g(y)$, under θ, is $V(\theta)$ and since $E_\theta\{g(y)\} = h(\theta)$, the lemma follows as in the derivation of the Rao-Cramer bound in the multiparameter case (Rao (1973), pp. 326–327).

The "if" part of Theorem 2.1 is immediate from Lemmas 2.1–2.3.

Remark that as conjectured by a referee, if the conditions of Theorem 2.1 fail to hold then $I(\theta)$ will cease to be positive definite only at $\beta = 0$. However, then in a neighbourhood of $\beta = 0$ as well, the convergence of maximum likelihood estimators should be slow and such estimators will perform poorly.

3. Poisson generalized linear mixed models

As opposed to the binary GLMM discussed in the last section, given $\gamma_1, \ldots, \gamma_K$, the observations y_{ij} $(1 \leq j \leq n_i, 1 \leq i \leq K)$ are now independent Poisson random variables with mean $a'(\theta^*_{ij}) = \exp(\theta^*_{ij})$, with $\theta^*_{ij} = x^T_{ij}\beta + z^T_i\gamma_i$. The assumption about the distribution of γ_i, however, remains the same. That is, as in the binary GLMM case, $\gamma_i \sim^{\text{iid}} N_q(0, \Sigma)$, with $\Sigma = (\sigma_{uv})$. This type of Poisson log-normal mixed model has been studied by many authors. For example, we refer to the recent study by Sutradhar and Qu (1998). These authors, however, have not examined the positive definiteness of the information matrix under such Poisson GLMMs.

As the purpose of this section is quite similar to that of the last section, we maintain the same notation as in the last section and state the main result as in the following theorem.

THEOREM 3.1. *Under the Poisson generalized linear mixed model, the information matrix $I(\theta)$ is p.d. for all θ if and only if both X and Z have full column rank.*

Here again, one can check that (2.4) holds. Hence the 'only if' part of Theorem 3.1 follows.

To prove the 'if' part of the theorem, we first assume that the conditions stated in Theorem 3.1 hold. Further let

$$U = [y_{11}(y_{11} - 1), \ldots, y_{K1}(y_{K1} - 1)]^T \tag{3.1}$$

under the Poisson model, whereas this U function was defined by (2.9) for the binary model.

Also define $g(y)$ as in (2.10) with U given by (3.1). The proof of the 'if' part can now be completed noting that Lemmas 2.1–2.3 hold even under the present setup. The proof of Lemma 2.2, with 'Theorem 2.1' in its statement replaced by 'Theorem 3.1' is given in Appendix B.

4. Concluding remarks

It is interesting to observe that the necessary and sufficient conditions for the binary and Poisson GLMMs are not identical. The conditions for the binary GLMM are more stringent than those for the Poisson GLMM. Note also that the conditions for the binary GLMM delete the contribution of $i \notin W$ only from Z but not from X. Thus the conditions for the positive definiteness of the Fisher information matrix $I(\theta)$ under a mixed generalized linear setup can vary from model to model and mere analogy with the usual linear model does not dictate the finer details underlying such conditions. For example, by analogy with the usual linear model, one might anticipate the necessity of both X and Z having full column rank for the positive definiteness of $I(\theta)$ (vide (2.4)). The sufficiency of this condition, however, is by no means obvious and, as Theorem 2.1 reveals, it is not sufficient for the binary GLMM. The point just noted, in addition to the algebraic complexities encountered in the proofs, makes the present problem nontrivial.

Before concluding, we briefly comment on some possible extensions of the present results. The first of these relates to the situation where the coefficient vector of γ_i depends on j in addition to i. In the linear regression case, this type of extended mixed model was studied by Laird and Ware (1982), among others. To deal with this extension in the present set up, we require to replace $z_i^T\gamma_i$ by $z_{ij}^T\gamma_i$ in the models considered in Sections 2 and 3. The following notation is then helpful. For $1 \le i \le K$ and $1 \le j \le s \le n_i$, let $\lambda_{ijs} = z_{ij}^T\Sigma z_{is}$; also, for any fixed i, let $\lambda_{(i)}$ be a vector with elements $\lambda_{ijs}(1 \le j \le s \le n_i)$. With the vector σ defined as before, then $\lambda_{(i)} = Z_i\sigma$, where the matrix Z_i, with elements dictated by those of $z_{ij}(1 \le j \le n_i)$, is easy to find. For example, if $z_{ij} = (z_{ij1}, \ldots, z_{ijq})^T$, then the first row of Z_i, which corresponds to $\lambda_{i11} = z_{i1}^T\Sigma z_{i1}$, is $(z_{i11}^2, 2z_{i11}z_{i12}, \ldots, 2z_{i11}z_{i1q}, z_{i12}^2, \ldots, 2z_{i12}z_{i1q}, \ldots, z_{i1q}^2)$. Let

$$\lambda = (\lambda_{(1)}^T, \ldots, \lambda_{(K)}^T)^T. \tag{4.1}$$

Then $\lambda = Z\sigma$, where

$$Z = (Z_1^T, \ldots, Z_K^T)^T. \tag{4.2}$$

In the above set up, it is not hard to see that for either the binary or the Poisson GLMM, the likelihood function depends on $\theta = (\beta^T, \sigma^T)^T$ only through $\psi = (\mu^T, \lambda^T)^T$, where μ is as before and λ is now given by (4.1). Hence, analogously to (2.4), for either model, in order to ensure the p.d.-ness of $I(\theta)$ for every θ, it is necessary that both X

and Z have full column rank, where X is as defined earlier and Z is now given by (4.2). For the Poisson model, this condition is also sufficient. One can prove this along the line of Lemmas 2.1–2.3 and Appendix B taking $g(y)$ as $g(y) = (y^T, U^T)^T$, where y is as before and U is now defined as a vector with elements $y_{ij}(y_{ij}-1)(1 \le j \le n_i, 1 \le i \le K)$ and $y_{ij}y_{is}(1 \le j < s \le n_i, 1 \le i \le K)$. While this completely settles the problem for the Poisson case, one encounters additional complexity, as in Section 2, for the binary GLMM. For the latter model, in the setup of the last paragraph too, equations like (2.6) and (2.7) hold. As such, analogously to the 'only if' part of Theorem 2.1, one gets a more stringent necessary condition for the p.d.-ness of $I(\theta)$, namely, $w > 0$ and both X and $\tilde{Z}$ have full column rank, where $\tilde{Z}$ is obtained by deleting from Z in (4.2) the submatrices Z_i for $i \notin W$, and W, w, and X are as defined earlier. This condition, however, may not be sufficient. More work is needed here and we hope that techniques similar to those in Lemmas 2.1–2.3 together with appendices may help in this regard.

Another possible extension of the present results concerns the situation where the dispersion matrix Σ is structured. Then the vector σ, defined earlier, takes a known functional form $\sigma = \sigma(\rho)$, where ρ is an unknown parametric vector of lower dimension than σ. In such a situation, the information matrix for $\theta^* = (\beta^T, \rho^T)^T$ is given by

$$I^*(\theta^*) = \{G(\rho)\}I(\theta)\{G(\rho)\}^T, \tag{4.3}$$

where $\theta = (\beta^T, \sigma^T)^T = (\beta^T, \sigma(\rho)^T)^T$, and $G(\rho) = \mathrm{diag}\{I, \partial\sigma(\rho)^T/\partial\rho\}$, with I being an identity matrix of order same as the dimension of β. Now suppose $\partial\sigma(\rho)^T/\partial\rho$ has full row rank for every ρ, a requirement which is met quite commonly (e.g. when Σ has an equi-correlation structure). Then by (4.3), the p.d.-ness of $I(\theta)$ for every θ implies that of $I^*(\theta^*)$ for every θ^*. Hence the conditions stated in Theorem 2.1 or 3.1 continue to remain sufficient for the p.d.-ness of $I^*(\theta^*)$ for all θ^*. This is satisfying since, with structured Σ, necessary and sufficient conditions for the p.d.-ness of $I^*(\theta^*)$ depend on the particular structure of Σ and hence cannot be specified in a unified manner.

Acknowledgements

We thank the referees for very constructive suggestions.

Appendix A

Proof of Lemma 2.2 for the binary generalized linear mixed model. Under the binary model, for $1 \le j \le n_i$, $1 \le i \le K$,

$$E_\theta(y_{ij}) = \pi_{ij} \tag{A.1}$$

where

$$\begin{aligned}\pi_{ij} &= \int_{-\infty}^{\infty} [\exp(x_{ij}^T\beta + z_i^T\gamma_i)/\{1+\exp(x_{ij}^T\beta + z_i^T\gamma_i)\}]\phi_q(\gamma_i;\Sigma)d\gamma_i, \\ &= \int_{-\infty}^{\infty} a_{ij}\phi(\tau_i)d\tau_i,\end{aligned}$$

with $a_{ij} = \exp(\mu_{ij} + \tau_i\sqrt{\lambda_i})/\{1+\exp(\mu_{ij} + \tau_i\sqrt{\lambda_i})\}$, and $\phi(\cdot)$ representing the standard univariate normal density. Similarly, for $1 \le i \le w$,

$$E_\theta(y_{i1}y_{i2}) = \delta_i, \tag{A.2}$$

where $\delta_i = \int_{-\infty}^{\infty} a_{i1}a_{i2}\phi(\tau_i)d\tau_i$. Now by (2.9), (2.10), (A.1) and (A.2),

$$h(\theta) = (\pi^T, \delta^T)^T, \tag{A.3}$$

where

$$\pi = (\pi_{11}, \ldots, \pi_{1n_1}, \ldots, \pi_{K1}, \ldots, \pi_{Kn_K})^T, \qquad \delta = (\delta_1, \ldots, \delta_w)^T.$$

In consideration of (A.1)–(A.3), $h(\theta)$ depends on θ only through $\psi = (\mu^T, \lambda^T)^T$ where μ and λ are defined earlier (cf. eqn (2.1)). It then follows that with Z, $\tilde{Z}$, λ, $\tilde{\lambda}$ and σ as defined earlier

$$\lambda = Z\sigma, \qquad \text{and} \qquad \tilde{\lambda} = (\lambda_1, \ldots, \lambda_w)^T = \tilde{Z}\sigma, \tag{A.4}$$

as $W = \{1, \ldots, w\}$. Under the condition of Theorem 2.1, the matrix $\tilde{Z}$ has full column rank. Hence by (A.4), $\sigma = (\tilde{Z}^T\tilde{Z})^{-1}\tilde{Z}^T\tilde{\lambda}$, that is $\lambda = Z(\tilde{Z}^T\tilde{Z})^{-1}\tilde{Z}^T\tilde{\lambda}$. Consequently, $h(\theta)$ depends on θ only through $\tilde{\psi} = (\mu^T, \tilde{\lambda}^T)^T$. Therefore, by (A.3),

$$\frac{\partial h(\theta)}{\partial \theta^T} = A(\tilde{\psi})\frac{\partial \tilde{\psi}}{\partial \theta^T} = A(\tilde{\psi})\begin{bmatrix} X & 0 \\ 0 & \tilde{Z} \end{bmatrix}, \tag{A.5}$$

where

$$A(\tilde{\psi}) = \begin{bmatrix} \dfrac{\partial \pi}{\partial \mu^T} & \dfrac{\partial \pi}{\partial \tilde{\lambda}^T} \\ \dfrac{\partial \delta}{\partial \mu^T} & \dfrac{\partial \delta}{\partial \tilde{\lambda}^T} \end{bmatrix}.$$

Under the condition stated in Theorem 2.1, the matrix in the extreme right of (A.5) has full column rank. It, therefore, remains to show that the matrix $A(\tilde{\psi})$ is nonsingular for every $\tilde{\psi} \equiv \tilde{\psi}(\theta)$.

By (A.1),

$$\frac{\partial \pi}{\partial \mu^T} = D(\tilde{\psi}), \tag{A.6}$$

where $D(\tilde{\psi})$ is a diagonal matrix of order $n \times n$ with $n = \sum_{i=1}^K n_i$, the (i,j)-th diagonal element being given by

$$d_{ij}(\tilde{\psi}) = \int_{-\infty}^{\infty} a_{ij}(1 - a_{ij})\phi(\tau_i)d\tau_i,$$

for $1 \le j \le n_i$, $1 \le i \le K$. Clearly, for every $\tilde{\psi}$, the quantities $d_{ij}(\tilde{\psi})$ are all positive. Hence $D(\tilde{\psi})$ is nonsingular and, in view of (A.5), it is enough to show that the matrix

$$A^*(\tilde{\psi}) = \left(\frac{\partial \delta}{\partial \tilde{\lambda}^T}\right) - \left(\frac{\partial \delta}{\partial \mu^T}\right)\{D(\tilde{\psi})\}^{-1}\left(\frac{\partial \pi}{\partial \tilde{\lambda}^T}\right) \tag{A.7}$$

is nonsingular for every $\tilde{\psi}$.

By (A.1), (A.2) and (A.3), δ does not involve μ_{ij} for any (i,j) that does not belong to the set $\{(1,1),(1,2),\ldots,(w,1),(w,2)\}$. Hence

$$\left(\frac{\partial \delta}{\partial \mu^T}\right)\{D(\tilde{\psi})\}^{-1}\left(\frac{\partial \pi}{\partial \tilde{\lambda}^T}\right) = \sum_{i=1}^{w}\sum_{j=1}^{2}\{d_{ij}(\tilde{\psi})\}^{-1}\left(\frac{\partial \delta}{\partial \mu_{ij}}\right)\left(\frac{\partial \pi_{ij}}{\partial \tilde{\lambda}^T}\right). \tag{A.8}$$

Again, for any i, j $(1 \le i \le w, j = 1, 2)$, by (A.1)–(A.3) all elements of the column vector $\partial\delta/\partial\mu_{ij}$ other than the i-th vanish, while by (A.1) and (A.4) all elements of the row vector $\partial\pi_{ij}/\partial\tilde{\lambda}^T$ other than the i-th vanish. Therefore, by (A.8)

$$\text{(A.9)} \qquad \left(\frac{\partial\delta}{\partial\mu^T}\right)\{D(\tilde{\psi})\}^{-1}\left(\frac{\partial\pi}{\partial\tilde{\lambda}^T}\right) = \operatorname{diag}(C_1(\tilde{\psi}), \ldots, C_w(\tilde{\psi})),$$

where for $1 \le i \le w$,

$$C_i(\tilde{\psi}) = \sum_{j=1}^{2}\{d_{ij}(\tilde{\psi})\}^{-1}\left(\frac{\partial\delta_i}{\partial\mu_{ij}}\right)\left(\frac{\partial\pi_{ij}}{\partial\lambda_i}\right).$$

From (A.1)–(A.4), it is also clear that for $1 \le i, i' \le w$, the quantity $\partial\delta_i/\partial\lambda_{i'}$ vanishes whenever $i \ne i'$. Hence by (A.7) and (A.9),

$$\text{(A.10)} \qquad A^*(\tilde{\psi}) = \operatorname{diag}[A_1(\tilde{\psi}), \ldots, A_w(\tilde{\psi})],$$

where for $1 \le i \le w$, $A_i(\tilde{\psi}) = \partial\delta_i/\partial\lambda_i - C_i(\tilde{\psi})$. Because of (A.10), it now suffices to show that for every $\tilde{\psi}$, the quantities $A_1(\tilde{\psi}), \ldots, A_w(\tilde{\psi})$ are all nonzero.

By (A.1) and (A.2), for $1 \le i \le w$,

$$\text{(A.11)} \qquad \frac{\partial\delta_i}{\partial\lambda_i} = \frac{1}{2\sqrt{\lambda_i}}\int_{-\infty}^{\infty} a_{i1}a_{i2}\{(1-a_{i1}) + (1-a_{i2})\}\tau_i\phi(\tau_i)d\tau_i.$$

Also, for $1 \le i \le w$, and $j = 1, 2$,

$$\text{(A.12)} \qquad \frac{\partial\delta_i}{\partial\mu_{ij}} = \int_{-\infty}^{\infty} a_{i1}a_{i2}(1-a_{ij})\phi(\tau_i)d\tau_i,$$

$$\text{(A.13)} \qquad \frac{\partial\pi_{ij}}{\partial\lambda_i} = \frac{1}{2\sqrt{\lambda_i}}\int_{-\infty}^{\infty} a_{ij}(1-a_{ij})\tau_i\phi(\tau_i)d\tau_i.$$

By (A.10)–(A.13), for $1 \le i \le w$,

$$\text{(A.14)} \qquad A_i(\tilde{\psi}) = B_{i1}(\tilde{\psi}) + B_{i2}(\tilde{\psi}),$$

where for $j = 1, 2$,

$$B_{ij}(\tilde{\psi}) = \frac{1}{2\sqrt{\lambda_i}}\left[\int_{-\infty}^{\infty} a_{i1}a_{i2}(1-a_{ij})\tau_i\phi(\tau_i)d\tau_i - \{d_{ij}(\tilde{\psi})\}^{-1}\right.$$
$$\left.\cdot\left\{\int_{-\infty}^{\infty} a_{i1}a_{i2}(1-a_{ij})\phi(\tau_i)d\tau_i\right\}\left\{\int_{-\infty}^{\infty} a_{ij}(1-a_{ij})\tau_i\phi(\tau_i)d\tau_i\right\}\right].$$

In view of (A.14), we shall now complete the proof by showing that for $1 \le i \le w$ and $j = 1, 2$, the quantities $B_{ij}(\tilde{\psi})$ are positive for every $\tilde{\psi}$.

To that effect, define for $1 \le i \le w$ and $j = 1, 2$,

$$\text{(A.15)} \qquad f_{ij}(\tau_i) = \{d_{ij}(\tilde{\psi})\}^{-1}a_{ij}(1-a_{ij})\phi(\tau_i),$$

and, recalling (A.6), note that $f_{ij}(\cdot)$ is a proper density over the real line. Also, define $m \equiv m(j)$ as $m(1) = 2$ and $m(2) = 1$. Then by (A.14), (A.15), for $1 \le i \le w$ and

$j = 1, 2,$

$$\text{(A.16)} \quad B_{ij}(\tilde{\psi}) = \frac{1}{2\sqrt{\lambda_i}} d_{ij}(\tilde{\psi}) \left[\int_{-\infty}^{\infty} a_{im} \tau_i f_{ij}(\tau_i) d\tau_i - \left\{ \int_{-\infty}^{\infty} a_{im} f_{ij}(\tau_i) d\tau_i \right\} \left\{ \int_{-\infty}^{\infty} \tau_i f_{ij}(\tau_i) d\tau_i \right\} \right]$$
$$= \frac{1}{2\sqrt{\lambda_i}} d_{ij}(\tilde{\psi}) \operatorname{cov}_{ij}(a_{im}, \tau_i),$$

where $\operatorname{cov}_{ij}(a_{im}, \tau_i)$ is the covariance between a_{im} and τ_i under the density $f_{ij}(\cdot)$ for τ_i. However, $f_{ij}(\cdot)$ has support over the entire real line and, by (A.1), both a_{i1} and a_{i2} are strictly increasing in τ_i. Therefore, $\operatorname{cov}_{ij}(a_{im}, \tau_i)$ is always positive and from (A.16) the proof is now complete.

Appendix B

Proof of Lemma 2.2 for the Poisson generalized linear mixed model. For the Poisson model, for every i, j,

$$E_\theta(y_{ij}) = E\{\exp(x_{ij}^T \beta + z_i^T \gamma_i)\} = \exp\left(x_{ij}^T \beta + \frac{1}{2} z_i^T \Sigma z_i\right) = \exp\left(\mu_{ij} + \frac{1}{2}\lambda_i\right),$$
$$E_\theta\{y_{ij}(y_{ij} - 1)\} = E\{\exp(2x_{ij}^T \beta + 2z_i^T \gamma_i)\} = \exp(2x_{ij}^T \beta + 2z_i^T \Sigma z_i) = \exp(2\mu_{ij} + 2\lambda_i).$$

Thus, in view of (3.1) and (2.10), $h(\theta) = E_\theta\{g(y)\}$ depends on θ only through $\psi = (\mu^T, \lambda^T)^T$, and we have

$$\text{(B.1)} \quad h(\theta) \equiv h^*(\psi) = (h_1^*(\psi)^T, h_2^*(\psi)^T)^T,$$

with

$$h_1^*(\psi) = \left[\exp\left(\mu_{11} + \frac{1}{2}\lambda_1\right), \ldots, \exp\left(\mu_{1n_1} + \frac{1}{2}\lambda_1\right), \ldots, \exp\left(\mu_{Kn_K} + \frac{1}{2}\lambda_K\right)\right]^T,$$
$$h_2^*(\psi) = [\exp(2\mu_{11} + 2\lambda_1), \ldots, \exp(2\mu_{K1} + 2\lambda_K)]^T.$$

By (2.2) and (B.1),

$$\text{(B.2)} \quad \frac{\partial h(\theta)}{\partial \theta^T} = \frac{\partial h^*(\psi)}{\partial \psi^T} \frac{\partial \psi}{\partial \theta^T} = \left(\frac{\partial h^*(\psi)}{\partial \psi^T}\right) \begin{pmatrix} X & 0 \\ 0 & Z \end{pmatrix}.$$

Under the condition stated in Theorem 3.1, the matrix in the extreme right of (B.2) has full column rank and it remains to show that for every $\psi \equiv \psi(\theta)$, the matrix $\frac{\partial h^*(\psi)}{\partial \psi^T}$ is nonsingular.

Now, by (B.1)

$$\text{(B.3)} \quad \frac{\partial h^*(\psi)}{\partial \psi^T} = \begin{bmatrix} A_{11}(\psi) & A_{12}(\psi) \\ A_{21}(\psi) & A_{22}(\psi) \end{bmatrix},$$

where

$$\begin{aligned} A_{11}(\psi) &= \frac{\partial h_1^*(\psi)}{\partial \mu^T} \\ &= \operatorname{diag}\Bigg[\exp\left(\mu_{11}+\frac{1}{2}\lambda_1\right),\ldots,\exp\left(\mu_{1n_1}+\frac{1}{2}\lambda_1\right),\ldots, \\ &\qquad \exp\left(\mu_{K1}+\frac{1}{2}\lambda_K\right),\ldots,\exp\left(\mu_{Kn_K}+\frac{1}{2}\lambda_K\right)\Bigg], \end{aligned}$$

and

$$A_{12}(\psi) = \frac{\partial h_1^*(\psi)}{\partial \lambda^T} = \frac{1}{2}A_{11}(\psi)E,$$

with

$$E = \begin{bmatrix} 1_{n_1} & 0 & \ldots & 0 \\ 0 & 1_{n_2} & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & 1_{n_K} \end{bmatrix},$$

1_u being the $u \times 1$ vector having all elements unity,

$$A_{21}(\psi) = \frac{\partial h_2^*(\psi)}{\partial \mu^T} = 2[\exp(2\mu_{11}+2\lambda_1)M_1,\ldots,\exp(2\mu_{K1}+2\lambda_K)M_K],$$

with M_i as a $K \times n_i$ matrix having 1 in the $(i,1)$-th position and 0 elsewhere,

$$A_{22}(\psi) = \frac{\partial h_2^*(\psi)}{\partial \lambda^T} = 2\operatorname{diag}[\exp(2\mu_{11}+2\lambda_1),\ldots,\exp(2\mu_{K1}+2\lambda_K)].$$

It is clear that $A_{11}(\psi)$ is always p.d.. Consequently, it follows from (B.3) that

$$\text{(B.4)}\quad \begin{aligned} \det\left(\frac{\partial h^*(\psi)}{\partial \psi^T}\right) &= [\det\{A_{11}(\psi)\}][\det\{A_{22}(\psi) - A_{21}(\psi)(A_{11}(\psi))^{-1}A_{12}(\psi)\}] \\ &= [\det\{A_{11}(\psi)\}]\left[\det\left\{A_{22}(\psi) - \frac{1}{2}A_{21}(\psi)E\right\}\right]. \end{aligned}$$

Now because

$$A_{21}(\psi)E = 2[\exp(2\mu_{11}+2\lambda_1)M_1 1_{n_1},\ldots,\exp(2\mu_{K1}+2\lambda_K)M_K 1_{n_K}] = A_{22}(\psi),$$

it follows from (B.4) that

$$\det\left(\frac{\partial h^*(\psi)}{\partial \psi^T}\right) = [\det\{A_{11}(\psi)\}]\left[\det\left\{\frac{1}{2}A_{22}(\psi)\right\}\right] > 0,$$

for every ψ. Thus the lemma is proved for the Poisson model.

References

Breslow, N. E. (1984). Extra-Poisson variation in log-linear models, *Appl. Statist.*, **33**, 38–44.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models, *J. Amer. Statist. Assoc.*, **88**, 9–25.

Breslow, N. E. and Lin, X. (1995). Bias correction in generalized linear models with a single component of dispersion, *Biometrika*, **82**, 81–92.

Jiang, J. (1998). Consistent estimators in generalized linear mixed models, *J. Amer. Statist. Assoc.*, **93**, 720–729.

Karim, M. R. and Zeger, S. L. (1992). Generalized linear models with random effects: Salamander mating revisited, *Biometrics*, **48**, 631–644.

Laird, N. M. and Ware, J. H. (1982). Random effects models for longitudinal data, *Biometrics*, **38**, 963–974.

Lawless, J. F. (1987). Negative binomial regression models, *Canad. J. Statist.*, **15**, 209–225.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models, *J. Roy. Statist. Soc. Ser. B*, **58**, 619–678.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed., Chapman and Hall, Cambridge, U. K.

Moore, D. F. (1986). Asymptotic properties of moment estimators for overdispersed counts and proportions, *Biometrika*, **73**, 583–588.

Morton, R. (1987). A generalized linear model with nested strata of extra-Poisson variation, *Biometrika*, **74**, 247–257.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models, *J. Roy. Statist. Soc. Ser. A*, **135**, 370–384.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed., Wiley, New York.

Stiratelli, R., Laird, N. M. and Ware, J. H. (1984). Random effects models for serial observations with binary responses, *Biometrics*, **40**, 961–971.

Sutradhar, B. C. and Qu, Z. (1998). On approximate likelihood inference in a Poisson mixed model, *Canad. J. Statist.*, **26**, 169–186.

Waclawiw, M. A. and Liang, K.-Y. (1993). Prediction of random effects in generalized linear model, *J. Amer. Statist. Assoc.*, **88**, 171–178.

Williams, D. A. (1982). Extra-binomial variation in logistic linear models, *Appl. Statist.*, **31**, 144–148.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects: A Gibbs sampling approach, *J. Amer. Statist. Assoc.*, **86**, 79–86.