Appl. Statist. (2006)
55, *Part* 2, *pp.* 139–185

Double hierarchical generalized linear models

Youngjo Lee

Seoul National University, Korea

and John A. Nelder

Imperial College London, UK

[*Read before* The Royal Statistical Society *on Wednesday, September 28th, 2005, the President*, Professor D. Holt, *in the Chair*]

Summary. We propose a class of double hierarchical generalized linear models in which random effects can be specified for both the mean and dispersion. Heteroscedasticity between clusters can be modelled by introducing random effects in the dispersion model, as is heterogeneity between clusters in the mean model. This class will, among other things, enable models with heavy-tailed distributions to be explored, providing robust estimation against outliers. The *h*-likelihood provides a unified framework for this new class of models and gives a single algorithm for fitting all members of the class. This algorithm does not require quadrature or prior probabilities.

Keywords: Generalized linear models; Heavy-tailed distribution; Hierarchical generalized linear models; Hierarchical likelihood; *h*-likelihood; Joint generalized linear models; Random-effect models; Restricted maximum likelihood estimator; Stochastic volatility models

1. Introduction

Our framework has two components. The first is the extended class of models, based on generalized linear models (GLMs) but allowing in addition the joint modelling of the mean and dispersion, and the introduction of random effects in their linear predictors. The second is the use of hierarchical (or *h*-)likelihood for making inferences from these models. The resulting algorithm is numerically efficient while giving statistically valid inferences.

Hierarchical generalized linear models (HGLMs) (Lee and Nelder, 1996) were originally developed from an initial synthesis of GLMs, random-effect models and structured dispersion models (Lee and Nelder, 2001a) and extended to include models for temporal and spatial correlations (Lee and Nelder, 2001b). Further extensions can be made by allowing random effects in various components in HGLMs. In this paper we investigate models with random effects in the dispersion model, which gives heavy-tailed models, allowing robust inference against outliers. Abrupt changes among repeated measures arising from the same subject can also be modelled by introducing random effects in the dispersion. We shall show how assumptions about skewness and kurtosis can be altered by using random effects. Many models can be unified and extended by the use of double hierarchical generalized linear models (DHGLMs); these include mixed linear models (Verbeke and Molenberghs, 2000), generalized linear mixed models (GLMMs) (Breslow and Clayton, 1993), HGLMs (Lee and Nelder, 1996), multilevel models (Goldstein, 1995), random-coefficient models (Longford, 1993), joint models of mean and dispersions

Address for correspondence: Youngjo Lee, Department of Statistics, College of Natural Sciences, Seoul National University, NS40, San56-1, Shin Lim-Dong, Kwan Ak-Ku, Seoul 151-747, Korea.
E-mail: youngjo@plaza.snu.ac.kr

 0035–9254/06/55139

(Nelder and Lee, 1991), splines (Wahba, 1990), semiparametric models (Zeger and Diggle, 1994), growth curve models (Zimmerman and Nunez-Anton, 2001), frailty models (Hougaard, 2000), heavy-tailed models (Lange *et al.*, 1989), autoregressive conditional heteroscedasticity (ARCH) models (Engel, 1995), generalized ARCH (GARCH) and stochastic volatility (SV) models (Harvey *et al.*, 1994) in finance data, etc.

In the synthesis of the inferential tools that are needed for these broad classes of model, the h-likelihood (Lee and Nelder, 1996) plays a key role and gives a statistically (Section 4.1) and numerically efficient algorithm (Section 5). This algorithm can be used throughout the extended class of models and requires neither prior distributions of parameters nor quadrature for integration. DHGLMs can be decomposed into a set of interlinked GLMs. Thus, a great variety of models can be generated, fitted and compared by using the GLM iterative weighted least squares (IWLS) procedures. We can change the link function, allow various types of term in the linear predictor and use model selection methods for adding or deleting terms, not only for inferences about the mean and dispersion but also about the platykurtosis etc. Furthermore various model assumptions can be checked by applying GLM model checking procedures to the component GLMs. Further extensions of the class will be discussed.

2. Double hierarchical generalized linear models

Suppose that, conditional on the pair of random effects (u, a), the response y satisfies

$$E(y|u,a) = \mu,$$
$$\text{var}(y|u,a) = \phi V(\mu),$$

where ϕ is the dispersion parameter and $V(\cdot)$ is the variance function. We allow the following.

(a) An extended GLM (EGLM) for μ: given u, the linear predictor for μ takes the GLM form

$$\begin{aligned}\eta &= g(\mu)\\ &= X\beta + Zv,\end{aligned} \tag{1}$$

where $g(\cdot)$ is the link function, X and Z are respectively $n \times p_1$ and $n \times s_1$ model matrices, $v = g_m(u)$, for some monotone function $g_m(\cdot)$, are the random effects and β are the fixed effects. Parameters λ for u have the GLM form

$$\begin{aligned}\xi_m &= h_m(\lambda)\\ &= G_m\gamma_m\end{aligned} \tag{2}$$

where $h_m(\cdot)$ is the link function, G_m is the $s_1 \times p_3$ model matrix and γ_m are fixed effects.

(b) An EGLM for ϕ: given a, the linear predictor for ϕ takes the GLM form

$$\begin{aligned}\xi &= h(\phi)\\ &= G\gamma + Fb,\end{aligned} \tag{3}$$

where $h(\cdot)$ is the link function, G and F are respectively $n \times p_2$ and $n \times s_2$ model matrices, $b = g_d(a)$, for some monotone function $g_d(\cdot)$, are the random effects and γ are the fixed effects. Parameters α for a have the GLM form with

$$\begin{aligned}\xi_d &= h_d(\alpha)\\ &= G_d\gamma_d,\end{aligned} \tag{4}$$

Table 1. Various submodels of DHGLMs

Model type	*Model for the mean*	*Model for the dispersion*
JGLM†, ARCH	GLM, null model	GLM
HGLM	EGLM	GLM, null model
Multivariate t-model, SV model	GLM, null model	EGLM
DHGLM	EGLM	EGLM

†The joint GLM of Section 5.2.

where $h_d(\cdot)$ is the link function, G_d is an $s_2 \times p_4$ model matrix and γ_d are fixed effects. Here the subscripts m and d stand for the mean and dispersion respectively.

The model matrices allow both categorical and continuous covariates. In addition to conjugate distributions, various distributions are allowed for random effects.

We use the term EGLM to allow random effects in a GLM; for example the EGLM (1) becomes a GLM if $v = 0$. The number of GLMs in equations (2) and (4) equals the number of random components in equations (1) and (3) respectively. HGLMs with fixed known ϕ, such as Poisson or binomial distributions for the μ-model, reduce to EGLMs. The general HGLM with unknown ϕ, for example with normal, gamma or inverse Gaussian distributions for the μ-model, has a GLM for ϕ. GLMMs are the HGLMs with normal random effects. DHGLMs allow random effects in the dispersion and become HGLMs of Lee and Nelder (2001a) if $b = 0$. Particular forms of DHGLMs have been used in various fields, and the four main types are summarized in Table 1. In this paper we call parameters (ϕ, λ, α) dispersion parameters. For simplicity of argument we consider models having random effects in ϕ only. Noh *et al.* (2005) have shown that by allowing random effects in the λ-component we can have parameter estimates that are insensitive to misspecification of the distribution for random effects.

2.1. Models with heavy-tailed distributions

Outliers have been observed in many physical and sociological phenomena (Medina *et al.*, 2001). Here the Gaussian assumption is vulnerable to outliers. To have robust estimation against such outliers, heavy-tailed models have often been suggested. In this paper the subscript i is reserved for independent sampling units (subjects) and the subscript j for repeated measurements on each unit. Consider a simple linear model

$$y_{ij} = X_{ij}\beta + e_{ij} \tag{5}$$

where $e_{ij} = \sigma_{ij} z_{ij}$, $z_{ij} \sim N(0, 1)$ and $\phi_{ij} = \sigma_{ij}^2$. Here the kurtosis of e_{ij} (y_{ij}) is given by

$$E(e_{ij}^4)/\mathrm{var}(e_{ij})^2 = 3\,E(\phi_{ij}^2)/E(\phi_{ij})^2 \geqslant 3,$$

where equality holds if and only if ϕ_{ij} are fixed constants. Thus, by introducing a random component in ϕ_{ij}

$$\log(\phi_{ij}) = \gamma + b_i$$

we can make the distribution of e_{ij} heavier tailed than that for z_{ij}. For example, if $a_i = \exp(b_i) \sim k/\chi_k^2$, the $e_i = (e_{i1}, \ldots, e_{im})^{\mathrm{T}}$ follows a multivariate t-distribution (Lange *et al.*, 1989) with

$$E(\phi_{ij}) = \mathrm{var}(y_{ij}) = k\exp(\gamma)/(k-2).$$

The multivariate t-model allows correlations by assuming correlations between the z_{ij}, which can be fitted by using a method that is similar to that in Section 5.4; for a recent review see Lindsey and Lindsey (2006). When $k=1$ it becomes a Cauchy distribution. This model allows an explicit form for the marginal distribution of e_{ij} but is restricted to a single random-effect model. Such a restriction is not necessary to produce heavy tails. If $\exp(b_i)$ follows a gamma or inverse gamma distribution with $E\{\exp(b_i)\}=1$ we have $E(\phi_{ij}) = \mathrm{var}(y_{ij}) = \exp(\gamma)$. If b_i is Gaussian, correlations can be easily introduced. The use of a multivariate t-model gives robust estimation, reducing to a Gaussian model at $k=\infty$. This is generally true for other distributions for b_i, which reduce to a Gaussian model at $\mathrm{var}(b_i)=0$. Tails become heavier with $1/k$ $(\mathrm{var}(b_i))$.

By introducing random effects v_i in the mean we can describe heteroscedasticities of means between clusters, whereas by introducing random effects b_i in the dispersion we can describe that of dispersions between clusters, which can in turn describe abrupt changes between repeated measures. However, contrary to v_i in the model for the mean, b_i in the dispersion does not necessarily introduce correlations. Because $\mathrm{cov}(y_{ij}, y_{il}) = 0$ for $j \neq l$ for independent z_{ij}, there are two ways of expressing correlation in DHGLMs: either by introducing a random effect in the mean linear predictor $X_{ij}\beta + v_i$ or by assuming correlations between the z_{ij}. The current DHGLMs adopt the former approach, and the multivariate t-models the latter.

It is possible to have a heavy-tailed distribution with current HGLMs, without introducing a random component in ϕ. Yun *et al.* (2004) showed that a heavy-tailed Pareto distribution can be generated via exponential–inverse gamma HGLMs, in which the random effects have variance $\mathrm{var}(u) = 1/(\alpha-1)$. The random effects have infinite variance when $0 < \alpha \leqslant 1$. Yun *et al.* (2004) could nevertheless fit the model to Internet service times with α in this interval and found that the estimated distribution indeed had a long tail.

2.2. Models for finance data

Consider a process y_t in time where

$$y_t = z_t\sqrt{\phi_t}$$

for z_t a standard normal variable and ϕ_t a function of the history of the process at time t. In financial models, the responses are often mean corrected to assume null means (e.g. Kim *et al.* (1998)). The simplest ARCH of order 1 (ARCH(1)) model (Engel, 1995) takes the form

$$\phi_t = \gamma_0^* + \gamma y_{t-1}^2.$$

This is a DGLM with $\mu=0$, $V(\mu)=1$, $h(\phi)=\phi$, $G_t=(1, y_{t-1}^2)$ and $b=0$ in equation (4). The ARCH(1) model can be extended to the GARCH model by assuming that

$$\phi_t = \gamma_0^* + \gamma_2\phi_{t-1} + \gamma_1 y_{t-1}^2,$$

which can be written as

$$\phi_t = \gamma_0 + b_t,$$

where $\gamma_0 = \gamma_0^*/(1-\rho)$, $\rho = \gamma_1 + \gamma_2$, $b_t = \phi_t - \gamma_0 = \rho b_{t-1} + r_t$ and $r_t = \gamma_1(y_{t-1}^2 - \phi_{t-1})$. The natural analogue of exponential GARCH models is

$$\xi_t = \log(\phi_t) = \gamma_0 + b_t,$$

where $b_t = \xi_t - \gamma_0$. Here the logarithm appears as the GLM link function for the dispersion ϕ. If $r_t \sim N(0, \alpha)$, i.e. $b_t = \rho b_{t-1} + r_t \sim \mathrm{AR}(1)$, this becomes the most popular SV model, originating with Harvey *et al.* (1994).

If we take positive-valued responses y^2 all these finance models become mean models. For example SV models become gamma HGLMs with temporal random effects (see Section 5.4), satisfying

$$E(y^2|b) = \phi,$$
$$\mathrm{var}(y^2|b) = 2\phi^2,$$

which is equivalent to assuming $y^2|b \sim \phi\chi^2_1$. Thus, the HGLM method can be used directly to fit these SV models.

2.3. Joint splines

Nonparametric analogues of the parametric modelling approach have been developed. For example, Zeger and Diggle (1994) proposed a semiparametric model for longitudinal data, where the covariate entered parametrically as $x_i\beta$ and the time effect entered nonparametrically as $v_i(t_i)$. Consider the model

$$y_i = x_i\beta + f_m(t_i) + e_i,$$

where the functional form of $f_m(\cdot)$ is unknown and $e_i \sim N(0, \phi_i)$, with $\phi_i = \phi$. Analogous to the approach of Green and Silverman (1994), page 12, natural cubic splines can be used to fit $f_m(t_i)$, by maximizing the penalized likelihood

$$-\tfrac{1}{2}(y-\mu)^{\mathrm{T}}(y-\mu)/\phi - \tfrac{1}{2}v^{\mathrm{T}}Kv/\lambda,$$

where λ is the smoothing parameter and $K = QR^{-1}Q^{\mathrm{T}}$. Let $h_j = t_{j+1} - t_j$; then Q is the $n \times (n-2)$ matrix with entries $Q_{i,j}$, for $i = 1, \ldots, n$ and $j = 1, \ldots, n-2$, given by

$$Q_{j,j} = 1/h_j,$$
$$Q_{j+1,j} = -1/h_j - 1/h_{j+1},$$
$$Q_{j+2,j} = 1/h_{j+1},$$

with the remaining being 0, and R is the $(n-2) \times (n-2)$ symmetric invertible matrix with elements $R_{i,j}$, given by

$$R_{j,j} = h_j + h_{j+1},$$
$$R_{j+1,j} = R_{j,j+1} = h_{j+1}/6,$$

and $R_{i,j} = 0$ for $|i-j| \geqslant 2$. This leads to an HGLM

$$y_i = x_i\beta + v_i + e_i,$$

where $v_i = v_i(t_i)$ is a random component with a singular precision matrix K/λ, depending on t_i. Here $\mathrm{rank}(K) = n-2$, and we can always find an $n \times (n-2)$ matrix L such that

$$L^{\mathrm{T}}KL = I_{n-2},$$

where I_{n-2} is the identity matrix of dimension $n-2$. Let $v = Lr$, giving

$$v^{\mathrm{T}}Kv = r^{\mathrm{T}}r.$$

Thus, the natural cubic splines for $f_m(t_i)$ can be obtained by fitting

$$y = x\beta + Lr + e,$$

where $r \sim N(0, \lambda I_{n-2})$ and $e \sim N(0, \phi I_n)$ (Lee and Nelder, 2001b): see Section 5.4. In discussing Robinson (1991), Speed pointed out that smoothing splines are random-effect estimators. There can be several alternative random-effect models leading to the same marginal model and the h-likelihood inferences from these are equivalent (Lee and Nelder, 2005a). An alternative random-effect model by taking $L = Q(Q^{\mathrm{T}}Q)^{-1}$ and $r \sim N(0, \lambda R)$ was studied by Verbyla *et al.* (1999) and another representation, using B-splines, by Brumback and Rice (1998). By changing K in the penalized likelihood we can fit other splines, e.g. the L-spline (Welham *et al.*, 2004). This also means that we can use different splines for fitting these curves.

Relative to the nonparametric modelling of the mean structure, nonparametric covariance modelling has received little attention. With DHGLMs we can consider semiparametric dispersions or heavy-tailed distributions with serial correlations etc. Consider heterogeneity in ϕ_i. Suppose that

$$\log(\phi_i) = x_i\gamma + f_d(t_i),$$

with an unknown functional form for $f_d(\cdot)$; this leads to a DHGLM for the component ϕ_i

$$\log(\phi_i) = x_i\gamma + b_i(t_i),$$

where $b_i(t_i)$ is a random component with a singular precision matrix K/α. This can be fitted similarly by using a model

$$\log(\phi) = x\gamma + La,$$

where $a \sim N(0, \alpha I_{n-2})$. In this paper we estimate the smoothing parameters (α, λ) by treating them as dispersion parameters. Rigby and Stasinopoulos (1996) have developed code to fit cubic splines for both the mean and dispersion (the MADAM algorithm); these can be also fitted as a DHGLM, as shown above. Note that it is not necessary to use the same splines for the mean and dispersion; for example, we can use the L-splines for the mean and the B-splines for the dispersion. The use of different splines implies the use of different Ls for the corresponding random effects.

2.4. Altering skewness and kurtosis

In GLMs the higher order cumulants of y are given by

$$\kappa_{r+1} = \kappa_2(\mathrm{d}\kappa_r/\mathrm{d}\mu), \qquad \text{for } r \geqslant 2, \tag{6}$$

where $\kappa_2 = \phi V(\mu)$ (Patil, 1963). With HGLMs y can have cumulants that are different in form from those for the $y|u$ component. For example, the negative binomial distribution (which is equivalent to a Poisson–gamma HGLM) can be shown to satisfy

$$\begin{aligned} \kappa_1 &= \mu_0, \\ \kappa_2 &= \mu_0 + \lambda\mu_0^2, \\ \kappa_{r+1} &= \kappa_2(\mathrm{d}\kappa_r/\mathrm{d}\mu) \qquad \text{for } r = 2, 3, \ldots; \end{aligned}$$

thus it still obeys the rules for GLM skewness and kurtosis (6), although it does not have the form of a one-parameter exponential family. Thus, random effects in the mean may provide different skewness and kurtosis from those for the $y|v$ component, though they may mimic similar patterns to those from a GLM family of given variance. By introducing random effects in the dispersion

we can produce models with different patterns from the GLM ones. Consider DHGLMs with the mean model having no random effects. Let κ_i^* be cumulants for the GLM family of $y|a$ and κ_i be those for y. Then, $\kappa_1 \equiv E(y_{ij}) = \kappa_1^* \equiv E(y_{ij}|b_i) = \mu$, $\kappa_2^* \equiv \text{var}(y_{ij}|b_i) = \phi\, V(\mu)$ and $\kappa_2 \equiv E\{\text{var}(y_{ij}|b_i)\} = E(\kappa_2^*) = E(\phi)\, V(\mu)$, so that

$$\begin{aligned}\kappa_3 &= E(\kappa_3^*)\\ &= E(\phi^2)\, V(\mu)\, \mathrm{d}V(\mu)/\mathrm{d}\mu\\ &\geqslant E(\phi)^2\, V(\mu)\, \mathrm{d}V(\mu)/\mathrm{d}\mu\\ &= \kappa_2(\mathrm{d}\kappa_2/\mathrm{d}\mu),\end{aligned}$$

and

$$\begin{aligned}\kappa_4 &= E(\kappa_4^*) + 3\,\text{var}(\phi)\, V(\mu)^2\\ &\geqslant E(\kappa_4^*)\\ &= \{E(\phi^2)/E(\phi)^2\}\,\kappa_2(\mathrm{d}\kappa_3/\mathrm{d}\mu)\\ &\geqslant \kappa_2(\mathrm{d}\kappa_3/\mathrm{d}\mu).\end{aligned}$$

Thus, higher order cumulants no longer have the pattern of those from a GLM family.

Consider now the model (5) but with $X_{ij}\beta + v_i$ for the mean. Here, even if all the random components are Gaussian, y_{ij} can still have a skewed distribution because

$$E\{y_{ij} - E(y_{ij})\}^3 = E(e_{ij}^3) + 3\,E(e_{ij}^2 v_i) + 3\,E(e_{ij} v_i^2) + E(v_i^3)$$

is non-zero if (b_i, v_i, z_{ij}) are correlated. When $v_i = 0$, $\kappa_3 = E(e_{ij}^3) \neq 0$ if z_{ij} and ϕ_{ij} (and hence b_i) are correlated. In DHGLMs we can produce various skewed distributions by taking non-constant variance functions.

3. Examples

We give several examples of data analysis, using DHGLMs. Here the Poisson–gamma–normal DHGLM has a Poisson distribution for the $y|(u, a)$ component, the gamma distribution for the u-component and the normal distribution for the a-component. The conjugate Poisson DHGLM (Section 5.5) is the Poisson–gamma–inverse gamma DHGLM. Note that the inverse gamma distribution appears as the conjugate of the gamma distribution. Similarly, the gamma DHGLM (Section 5.5) is the gamma–inverse gamma–inverse gamma DHGLM.

3.1. Crack growth data

Hudak *et al*. (1978) presented some crack growth data, which are listed in Lu and Meeker (1993). There are 21 metallic specimens, each subjected to 120 000 loading cycles, with the crack lengths recorded every 10^4 cycles. We take t to be the number of cycles divided by 10^6, so $t_j = j/100$ for $j = 1, \ldots, 12$. The crack increment sequences look rather irregular. Let l_{ij} be the crack length of the ith specimen at the jth observation and let $y_{ij} = l_{ij} - l_{ij-1}$ be the corresponding increment of crack length, which always has a positive value. Models that describe the process of deterioration or degradation of units or systems are of interest and are also a key ingredient in processes that model failure events. Lu and Meeker (1993) and Robinson and Crowder (2000) proposed non-linear models with normal errors. Lawless and Crowder (2004) proposed to use a gamma process with independent increments but in discrete time. Their model is similar to the conjugate gamma HGLM (7), but using the covariate t_j. We found that the total crack size is a better covariate for the crack growth, so that the resulting model has non-independent increments.

From the normal probability plot in Fig. 1(a) for the HGLM (7) without a random component in the dispersion we can see the presence of outliers, caused by abrupt changes between repeated measures. Our final model is a conjugate gamma DHGLM with $V_m(u)=u^2$ and $V_d(a)=a^2$; the parameter estimates followed by their standard error estimates in parentheses are given by

$$\left.\begin{aligned}\eta_{ij}=\log(\mu_{ij})=\beta_0+l_{ij-1}\beta_l+v_i,\\ \xi_m=\log(\lambda)=\gamma_m,\\ \xi_{ij}=\log(\phi_{ij})=\gamma_0+t_j\gamma_t+b_i,\\ \xi_{di}=\log(\alpha_i)=\gamma_d.\end{aligned}\right\} \tag{7}$$

Now we want to test a hypothesis $H_0:\mathrm{var}(b_i)=0$ (i.e. the absence of random effects in the dispersion). Such a hypothesis is on the boundary of the parameter space, so the critical value for a deviance test is $\chi^2_{1,0.1}=2.71$ for a size 0.05 test (Chernoff, 1954). Here the difference in deviance $(-2\,p_{v,b,\beta}(h))$ is 14.95; thus a heavy tail is indeed necessary. From Fig. 1(b) for the DHGLM we see that most of the outliers, caused by abrupt changes between repeated measures, disappear when we introduce random effects into the dispersion.

Table 2 shows the results from the DHGLM and submodels. In this data set the regression estimators β are insensitive to the dispersion modelling. The DHGLM has the smallest standard error estimates, reflecting the information gain that is obtained by having proper dispersion modelling.

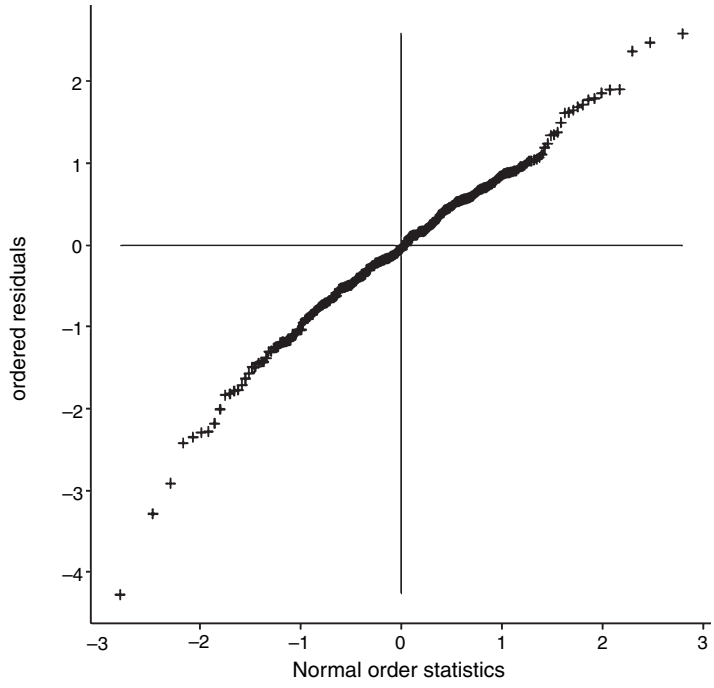
3.2. Data on epileptics

Thall and Vail (1990) presented longitudinal data from a clinical trial of 59 epileptics, who were randomized to a new drug or a placebo ($T=0$ or $T=1$). Base-line data that were available at the start of the trial included the logarithm of the average number of epileptic seizures recorded in the 8-week period preceding the trial (B), the logarithm of age (A) and visit (V: a linear trend, coded $(-3,-1,1,3)/10$). A multivariate response variable consisted of the seizure counts during 2-week periods before each of four visits to the clinic. Either random effects or extra-Poisson variation ($\phi>1$) could explain the overdispersion among repeated measures within a subject. Thall and Vail (1990), Breslow and Clayton (1993), Diggle *et al.* (1994) and Lee and Nelder (1996) have analysed these data by using various Poisson HGLMs. Lee and Nelder (2000) showed that both explanations are necessary to give an adequate fit to the data. Using residual plots they showed their final model to be better than the other models that they considered. However, those plots still showed apparent outliers, as shown in Fig. 2(a). Consider a model in the form of a conjugate Poisson DHGLM ($V_m(u)=u$ and $V_d(a)=a^2$) as follows:

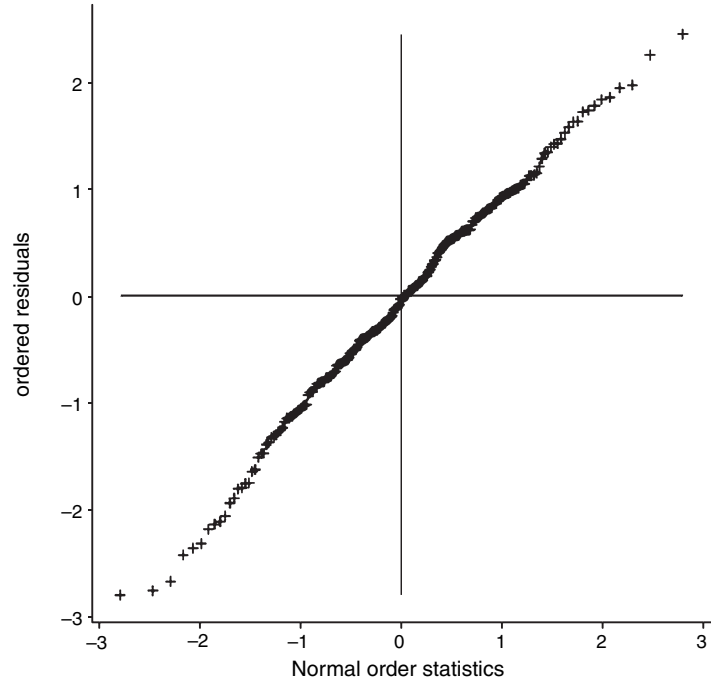
$$\begin{aligned}\eta_{ij}=\beta_0+T\beta_T+B\beta_B+T{*}B\beta_{T*B}+A\beta_A+V\beta_V+v_i,\\ \xi_{mi}=\log(\lambda_i)=\gamma_m,\\ \xi_{ij}=\log(\phi_{ij})=\gamma_0+B\gamma_B+b_i,\\ \xi_{di}=\log(\alpha_i)=\gamma_d.\end{aligned}$$

The difference in deviance for the absence of a random component $\alpha=\mathrm{var}(b_i)=0$ between a DHGLM and an HGLM with structured dispersion is 113.52, so the component b_i is necessary. From the normal probability plot for the DHGLM (Fig. 2(b)) we see that an apparent outlier vanishes.

Table 3 shows the results from the DHGLM and submodels. The regression estimator β_{T*B} is not significant at the 5% significance level under an HGLM and quasi-HGLM, whereas it is significant under an HGLM with structured dispersion and a DHGLM. Thus, proper dispersion



(a)



(b)

Fig. 1. Normal probability plots for (a) the HGLM and (b) the DHGLM of the crack growth data

Table 2. Summaries of analysis for the crack growth data

Parameter	*HGLM*		*HGLMSD*†		*DHGLM*	
	Estimate	*Standard error*	*Estimate*	*Standard error*	*Estimate*	*Standard error*
β_0	−5.63	0.09	−5.66	0.09	−5.62	0.08
β_l	2.38	0.07	2.41	0.07	2.38	0.06
γ_0	−3.32	0.10	−2.72	0.21	−4.08	0.21
γ_t			−10.58	2.93	−10.38	2.96
$\log(\lambda)$	−3.40	0.35	−3.42	0.35	−3.38	0.34
$\log(\alpha)$					−1.45	0.40
$-2\,p_{v,b,\beta}(h)$	−1509.3		−1522.0		−1536.9	

†HGLM with structured dispersion.

modelling gives a more powerful test. Anyhow the DHGLM has the smallest standard error estimates, reflecting the gain in information from having proper dispersion modelling.

3.3. Pound–dollar exchange data

We analyse daily observations of the weekday closing exchange rates for the UK pound sterling–US dollar from October 1st, 1981, to June 28th, 1985. We follow Harvey *et al.* (1994) in using, as the response, the 936 mean-corrected returns

$$y_t = 100\left\{\log(r_t/r_{t-1}) - \sum \frac{\log(r_i/r_{i-1})}{n}\right\},$$

where r_t denotes the exchange rate at time t. Harvey *et al.* (1994), Shephard and Pitt (1997), Kim *et al.* (1998) and Durbin and Koopman (2000) fitted the SV model

$$\log(\phi_t) = \gamma_0 + b_t, \tag{8}$$

where $b_t = \rho b_{t-1} + r_t \sim \mathrm{AR}(1)$ with $r_t \sim N(0, \alpha)$. The efficiency of the estimator of Harvey *et al.* (1994) was improved by Shephard and Pitt (1997) by using a Markov chain Monte Carlo (MCMC) method, and this was again improved in speed by Kim *et al.* (1998). Durbin and Koopman (2000) developed an importance sampling method for both the maximum likelihood (ML) and the Bayesian procedures. The DHGLM estimates are

$$\log(\phi_t) = -0.874(0.200) + b_t,$$
$$\log(\alpha) = -3.515(0.278).$$

Table 4 shows the results. The parameterization

$$\sigma_t = \sqrt{\phi_t}$$
$$= \kappa \exp(b_t/2), \qquad \kappa = \exp(\gamma_0/2),$$

has a clearer economic interpretation (Kim *et al.*, 1998). SV models have attracted much attention recently as a way of allowing clustered volatility in asset returns. However, despite their intuitive appeal these models have been used less frequently than ARCH–GARCH models in

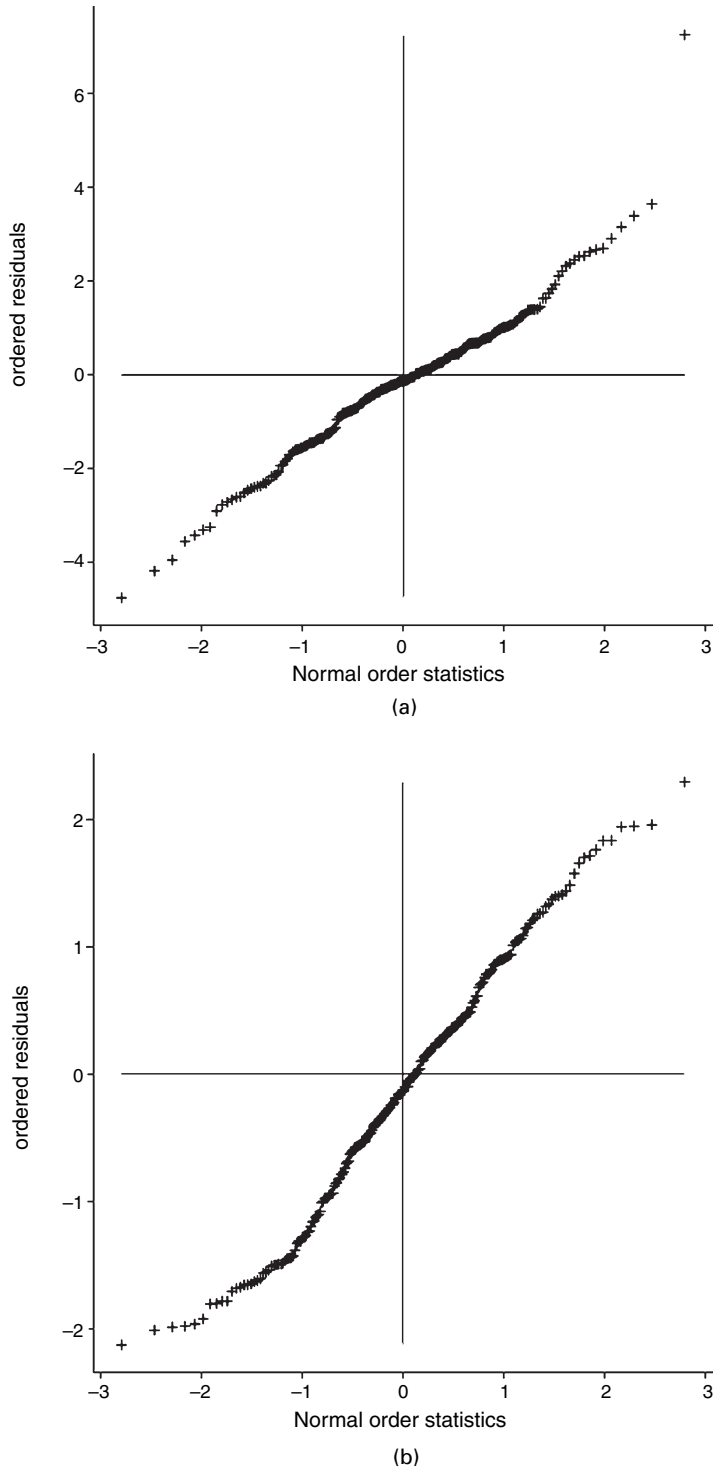


Fig. 2. Normal probability plots for (a) the HGLM and (b) the DHGLM of the epileptics data

Table 3. Summaries of analyses for the epileptics data

Parameter	*HGLM*		*HGLMQ*†		*HGLMSD*‡		*DHGLM*	
	Estimate	*Standard error*	*Estimate*	*Standard error*	*Estimate*	*Standard error*	*Estimate*	*Standard error*
β_0	−1.38	1.15	−1.66	1.13	−1.37	1.11	−1.43	0.91
β_B	0.88	0.12	0.91	0.12	0.88	0.12	0.91	0.10
β_T	−0.89	0.37	−0.92	0.39	−0.91	0.36	−0.77	0.30
β_{T*B}	0.34	0.19	0.36	0.19	0.36	0.18	0.31	0.15
β_A	0.52	0.34	0.58	0.33	0.51	0.28	0.50	0.27
β_V	−0.29	0.10	−0.29	0.15	−0.28	0.17	−0.28	0.12
γ_0			0.83	0.11	−0.05	0.28	−0.62	0.27
γ_B					0.48	0.14	0.37	0.14
$\log(\lambda)$	−1.47	0.22	−1.82	0.31	−1.80	0.30	−2.12	0.29
$\log(\alpha)$							−0.34	0.21
$-2\,p_{v,b,\beta}(h)$	1346.2		1263.6		1255.4		1141.9	

†Quasi-HGLM.
‡HGLM with structured dispersion.

Table 4. Summaries of analyses for the daily exchange rate data†

	Kim et al. (1998) estimate	*Shephard and Pitt (1997) estimate*	*Durbin and Koopman (2000) estimate*	*DHGLM estimate*	*LI*	*UI*
$\exp(\gamma_0/2)$	0.649	0.659	0.634	0.646	0.533	0.783
$\sqrt{\alpha}$	0.158	0.138	0.172	0.172	0.131	0.226

†LI and UI stand for the 95% lower and upper confidence bounds of the DHGLM fit.

applications. This is partly due to the difficulty that is associated with estimation in SV models (Shephard, 1996), where the use of marginal likelihood involves intractable integration, the integral being n dimensional (total sample size). Thus, computationally extensive methods such as Bayesian MCMC and simulated EM algorithms have been developed. The two previous analyses based on a Bayesian MCMC method report on the $\exp(\gamma_0/2)$ and $\sqrt{\alpha}$-scales, whereas with our DHGLM procedure we report on the γ_0- and $\log(\alpha)$ scales. For comparisons we use a common scale. The MCMC method assumes priors whereas the likelihood method does not, so the results may not be directly comparable. Note first that the two Bayesian MCMC analyses by Kim *et al*. (1998) and Shephard and Pitt (1997) are similar, and that the two likelihood analyses, by our h-likelihood method and Durbin and Koopman's (2000) importance sampling method, are also similar. Table 4 shows that the 95% confidence bounds for our method contain all the other estimates.

With this example we compare the h-likelihood estimates with other existing estimators for finance data. Parameter estimates can be directly computed without resorting to computationally intensive simulation methods such as importance sampling or Monte Carlo methods. An advantage of our method is a direct computation of standard error estimates from the Hessian matrix. Furthermore, these models for finance data can now be extended in various ways, allowing mean drift, non-constant variance functions, etc. With other data having additional

covariates such as days, weeks and months we have found that weekly or monthly random effects are useful for modelling ϕ_t, so further studies of these new models for finance data would be interesting.

3.4. Joint cubic splines

Suppose that the data are generated from a model, for $i = 1, \ldots, 100$,

$$y_i = f_m(x_i) + z_i \sqrt{f_d(x_i)},$$

where $z_i \sim N(0, 1)$; following Wahba (1990), page 45, we assume that

$$f_m(x_i) = \exp[4.26\{\exp(x_i) - 4\exp(-2x_i) + 3\exp(-3x_i)\}]$$

and we take

$$f_d(x_i) = 0.07\exp\{-(x_i - \bar{x})^2\}.$$

Because the actual functional forms of $f_m(\cdot)$ and $f_d(x_i)$ are assumed unknown, we applied joint splines using the natural cubic splines of Section 2.3. For fitting the mean and dispersion we use the normal–normal–normal DHGLM of Section 2.3, replacing t_i with x_i, giving

$$\mu_i = \beta_0 + x_i\beta_1 + v_i(x_i),$$
$$\log(\phi_i) = \gamma_0 + x_i\gamma_1 + b_i(x_i).$$

With 100 observations the fitted splines for the mean and variance, which are shown in Fig. 3, are satisfactory. With DHGLMs extensions to data in the form of counts or proportions are immediate.

4. *h*-likelihood

Lee and Nelder (1996) proposed to use *h*-likelihood for inferences about models having unobserved random effects. Its usefulness had not been fully appreciated because the worst result,

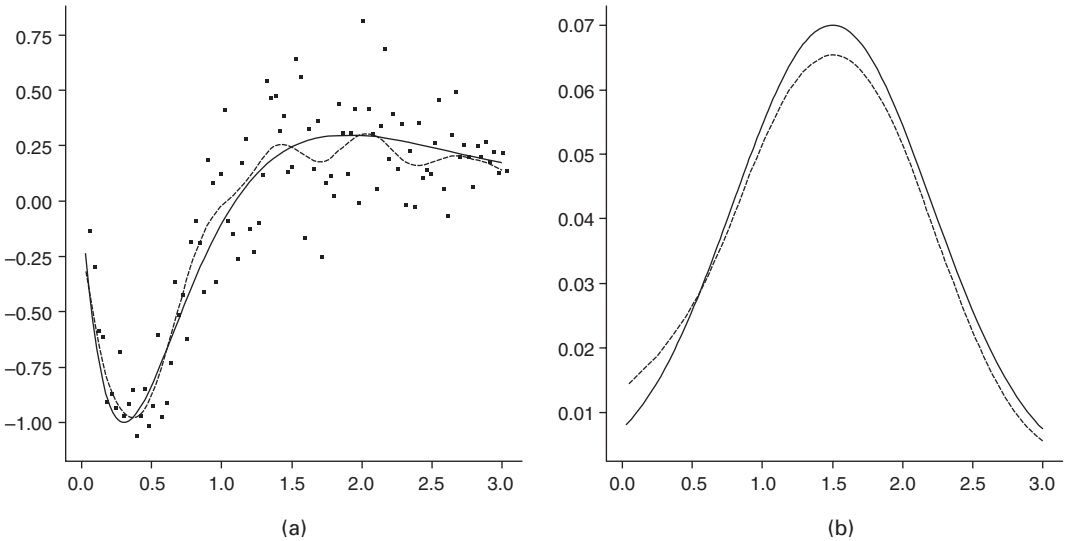


Fig. 3. Cubic splines for (a) the mean and (b) the variance (———, true value; -------, estimates)

which came from matched binary data, was held to characterize the entire HGLM estimation scheme. The fact that the method produced sensible estimators in other cases was ignored. Even for the analysis of binary data we have found no procedure (including MCMC-type procedures) that works better than the *h*-likelihood procedure (Noh and Lee, 2004) either in bias or mean-square error with small or large samples. In our view the *h*-likelihood method has been unreasonably criticized, the criticisms deriving from misunderstanding of the details of its specification.

One criticism was that it apparently provided non-invariant inferences for trivial re-expressions of the underlying model. Lee and Nelder (2005b) showed how defining the *h*-likelihood on the right scale avoids such difficulties. Another criticism has been about the statistical efficiency in the analysis of binary data. Much of this criticism stems from a confusion with other methods using joint maximization to obtain all the parameter estimates or some other method using *ad hoc* approximations. The restricted maximum likelihood (REML) method requires different functions to be maximized to estimate mean and dispersion parameters. Our generalization of REML shows that appropriate adjusted profile *h*-likelihoods must be used for the estimation of fixed (non-random) parameters (Lee and Nelder, 1996, 2001a). The remaining criticisms resulted from confusion with other methods which miss some important terms that are present in the *h*-likelihood procedure.

For inferences from DHGLMs, we propose to use an *h*-likelihood in the form

$$h = \log\{f(y|v,b;\beta,\phi)\} + \log\{f(v;\lambda)\} + \log\{f(b;\alpha)\},$$

where $f(y|v,b;\beta,\phi)$, $f(v;\lambda)$ and $f(b;\alpha)$ denote the conditional density functions of y given (v,b), and those of v and b respectively. In forming the *h*-likelihood we use the scales of (v,b) to be those on which the random effects occur linearly in the linear predictors: see Lee and Nelder (2005b) on why this scale is important in giving meaningful inferences. Bjørnstad (1996) showed that a likelihood such as the *h*-likelihood carries all the information in the data about the unobserved quantities (v,b) and the fixed parameters. Pawitan (2001) has given a further justification for the *h*-likelihood (called by him extended likelihood) as a predictive likelihood for inferences about future observations. The marginal (log-)likelihood $L_{v,b}$ can be obtained from h via an integration:

$$\begin{aligned} L_{v,b} &= \log\left\{\int \exp(h)\,\mathrm{d}v\,\mathrm{d}b\right\} \\ &= \log\left\{\int \exp(L_v)\,\mathrm{d}b\right\} \\ &= \log\left\{\int \exp(L_b)\,\mathrm{d}v\right\}, \end{aligned}$$

where $L_v = \log\{\int \exp(h)\,\mathrm{d}v\}$ and $L_b = \log\{\int \exp(h)\,\mathrm{d}b\}$. The marginal likelihood $L_{v,b}$ provides legitimate inferences about the fixed parameters. However, for general inferences it is not enough, because it is totally uninformative about the unobserved random parameters (v,b).

4.1. *h*-likelihood procedures and their statistical efficiency

There has been a widespread belief that the *h*-likelihood gives severely biased estimation, especially with binary data. If the number of nuisance parameters increases with the sample size it is important to use an adjusted profile likelihood, eliminating the nuisance parameters, to estimate the remaining parameters. It is especially important with random-effect models where the number of nuisance random effects often grows with the sample size.

Table 5. Criteria for effects in HGLMs

Criterion	*Arguments*	*Estimated*	*Eliminated*	*Approximation*
In general				
h	$v, \beta, \gamma, \gamma_m$	v, β	None	
$p_\beta(L_v)$	γ, γ_m	γ, γ_m	v, β	$p_{v,\beta}(h)$
For binary data				
h	$v, \beta, \gamma, \gamma_m$	v	None	h
L_v	β, γ, γ_m	β	v	$p_v(h)$
$p_\beta(L_v)$	γ, γ_m	γ, γ_m	v, β	$p_{v,\beta}(h)$

Let l be a likelihood (either a marginal likelihood L or an h-likelihood h) with nuisance effects δ. For HGLMs Lee and Nelder (2001a) considered a class of *adjusted profiling* functions $p_\delta(l)$, defined by

$$p_\delta(l) = \left(l - \frac{\log[\det\{D(l,\delta)/2\pi\}]}{2} \right) \Bigg|_{\delta=\tilde{\delta}}$$

where $D(l,\delta) = -\partial^2 l/\partial\delta^2$ and $\tilde{\delta}$ solves $\partial l/\partial\delta = 0$. For eliminating β the use of $p_\beta(L_v)$ is equivalent to the first-order approximation to the conditional likelihood that is obtained by conditioning on $\tilde{\beta}$ under parameter orthogonality (Cox and Reid, 1987), whereas for v the use of $p_v(h)$ is equivalent to integrating them out by using the first-order Laplace approximation. In mixed linear models $p_\beta(L_v)$ is called the residual or restricted likelihood (Harville, 1977). Table 5 summarizes estimation criteria for HGLMs.

We propose to use the h-likelihood h for inferences about v, the marginal likelihood L_v for β and the restricted likelihood $p_\beta(L_v)$ for the dispersion parameters. In general $p_{\beta,v}(h)$ is approximately $p_\beta\{p_v(h)\}$ and therefore $p_\beta(L_v)$. When L_v is numerically difficult to obtain, we propose to use $p_v(h)$ and $p_{\beta,v}(h)$ as approximations to L_v and $p_\beta(L_v)$. Lee and Nelder (1996) found that the deviance differences that are constructed from h and $p_v(h)$ are often very similar, so the use of h for estimating β often provides satisfactory estimation. This method works generally well in various models: see the simulation studies of Poisson and binomial models with moderately large binomial denominator (Lee and Nelder, 2001a), of frailty models (Ha *et al.*, 2001) and of mixed linear models with censoring (Ha *et al.*, 2002). However, for the analysis of binary data $p_v(h)$ should be used to eliminate an undesirable bias in the estimation of β (Yun and Lee, 2004a). The second-order Laplace approximation improves the accuracy of the estimators for dispersion parameters (Noh and Lee, 2004), and this should be used for non-normal random effects (Lee and Nelder, 2001a).

Breslow and Clayton (1993) introduced penalized quasi-likelihood (PQL) estimators, this being a generalization of the REML procedure for mixed linear models, and equivalent to Schall's (1991) estimator. The method is based on a quadratic approximation to the h-likelihood and they therefore proposed to use it only for the limited class of models such as GLMMs with a fixed known ϕ, i.e. EGLMs with normal random effects in Section 2. However, in this limited class of models such as binomial and Poisson GLMMs the PQL method can produce a serious bias in parameter estimators. Breslow and Lin (1995) derived correction factors which remove the asymptotic bias of the PQL estimators by using a Taylor series expansion for marginal likelihood about $v = 0$. However, their corrected PQL (CPQL) estimators still fail seriously for large values of the variance components (Lin and Breslow, 1996).

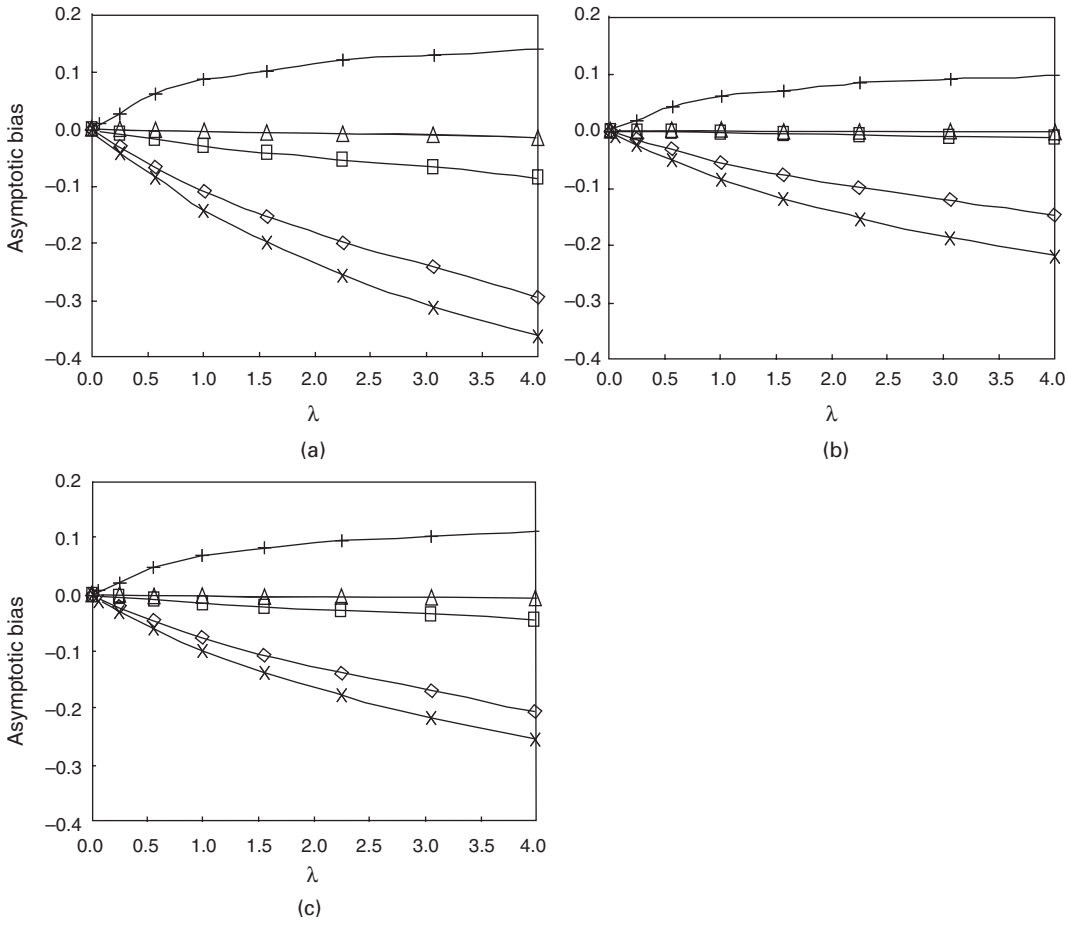


Fig. 4. Asymptotic bias in estimating β_1 (×, PQL; +, CPQL; ◇, $H(0)$; □, $H(1)$; △, $H(2)$): (a) $J=2$, $\beta_2=0$; (b) $J=3$, $\beta_2=0$; (c) $J=2$, $\beta_2=1$

Consider a Bernoulli model

$$\log\{p_{ij}/(1-p_{ij})\}=\beta_0+\beta_1(j-1)+\beta_2 x_i+v_i$$

where $p_{ij}=P(y_{ij}=1|v_i)$, $\beta_0=\beta_1=1$, x_i is 0 for the first half of individuals and 1 otherwise and $v_i\sim N(0,\lambda)$. First consider the model without x_i, i.e. with $\beta_2=0$. When $J=2$ this model is that for binary matched pairs. Following Noh and Lee (2004) we compute asymptotic biases of proposed estimators. Fig. 4(a) shows the asymptotic biases of the PQL, CPQL and the h-likelihood methods. In h-likelihood methods, $H(0)$ uses the criterion h for β, $H(1)$ uses the criterion $p_v(h)$ for β in Table 5 for the binary data and $H(2)$ uses the second-order approximation in estimating λ (Lee and Nelder, 2001a). PQL ignores terms such as $\partial\tilde{v}/\partial\theta$ in $H(0)$ and this causes an additional bias (Lee and Nelder, 2001a). By comparing $H(0)$ and $H(1)$ we see the bias caused by not using $p_v(h)$ in estimating β for binary data. A comparison of $H(1)$ and CPQL shows that the h-likelihood procedure is better. The second-order correction of CPQL is bad (Lin and Breslow, 1996). The bias in $H(1)$ is caused by biased estimation for λ (Yun and Lee, 2004a), so by using the second-order approximation the asymptotic bias can be greatly reduced. Furthermore such a bias can also be totally eliminated (Noh and Lee, 2004). When $J=2$, the

maximum biases of $H(1)$ and $H(2)$ at $\lambda = 4$ are respectively -0.087 and -0.014. When $J = 3$ in Fig. 4(b), the maximum biases of $H(1)$ and $H(2)$ are -0.011 and -0.0019 respectively. Lee (2001) noted that biases of the h-likelihood method would be caused by the fact that, regardless of how many distinct b_is are realized in the model, there are only a few distinct values for $\hat{b}_i$. In binary matched pairs there are only three distinct values for $\hat{v}_i$ because they depend only on $\Sigma_j y_{ij}$, which can be $0, 1, 2$. Now consider the model with $\beta_2 = 1$. In this model there are now six distinct values for $\hat{v}_i$ because $\Sigma_j p_{ij}$ has two distinct values depending on the value of x_i. In Fig. 4(c) the maximum biases of $H(1)$ and $H(2)$ are reduced to -0.044 and -0.0060 respectively. Thus, these maximum asymptotic biases decrease rapidly with either additional covariates or an increase in J. In multicomponent models, $H(2)$ can be computationally slow. If several covariates are available, $H(1)$ can maintain nominal confidence levels for accurate inferences (Noh *et al.*, 2004). They also showed that for very large binary data sets it is the only method that is feasible in practice (because of computation times) that maintains the nominal level of confidence. With only one or no covariate it is easy to eliminate totally the bias to maintain statistical efficiency (Noh and Lee, 2004). In small samples, e.g. $n \leqslant 50$, $H(0)$ works best among the methods that are considered (including the marginal ML estimator). Thus, regardless of whether samples are small or large in binary data the h-likelihood always give statistically sensible estimators.

Estimation procedures for HGLMs can be easily extended to DHGLMs as shown in Section 5.5. Some simulation studies are available for normal DHGLMs (Yun and Lee, 2004b) and for binary DHGLMs (Noh *et al.*, 2005). These show that h-likelihood gives numerically efficient estimation and that the resulting estimators are sufficiently accurate to be as good as or better than other available methods for these classes of models. Another advantage is that the standard errors are easily estimated from the second derivatives of likelihoods, so resulting interval estimators maintain the nominal levels.

5. Fitting methods for generalized linear model type classes of models

To develop fitting and inferential tools for DHGLMs three procedures are essential:

(a) the fitting of two interconnected GLMs for the mean and dispersion,
(b) the fitting of an augmented GLM for the random effects and
(c) the fitting of linear transformations for correlated random effects.

DHGLMs can be decomposed into an interconnected set of component GLMs and thus can be fitted by component (augmented or not) GLMs with or without linear transformations for correlation. Thus, GLMs serve as basic building-blocks to define and fit the new class of extended models. The GLM attributes of a DHGLM are summarized in Table 6, which shows the overall structure of the extended models. We define components as either fixed unknown constants (parameters) or unobservables (random effects). In our framework each component has its own GLM, so the development of inferential procedures for the components is straightforward. For example, if we are interested in model checking for the component ϕ (γ) in Table 6 we can use the procedures for a GLM having response d^*. In this Section we first review the GLM class of models with its associated fitting methods, and we then show how these three procedures can be combined to fit DHGLMs. Components of various GLM-type classes of models are summarized in Table 7.

5.1. Generalized linear models and iterative weighted least squares

Suppose that the responses y satisfy

Table 6. GLM attributes for DHGLMs†

	EGLM for the mean		*EGLM for the dispersion*	
	Augmented GLM	*GLM*	*Augmented GLM*	*GLM*
	β *(fixed)*		γ *(fixed)*	
Response	y		d^*	
Mean	μ		ϕ	
Variance	$\phi\, V(\mu)$		$2\phi^2$	
Link	$\eta = g(\mu)$		$\xi = h(\phi)$	
Linear predictor	$X\beta + Zv$		$G\gamma + Fb$	
Deviance component	d		gamma(d^*, ϕ)	
Prior weight	$1/\phi$		$(1-q)/2$	
	u *(random)*	λ *(fixed)*	a *(random)*	α *(fixed)*
Response	ψ_m	d_m^*	ψ_d	d_d^*
Mean	u	λ	a	α
Variance	$\lambda\, V_m(u)$	$2\lambda^2$	$\alpha\, V_d(a)$	$2\alpha^2$
Link	$\eta_m = g_m(u)$	$\xi_m = h_m(\lambda)$	$\eta_d = g_d(a)$	$\xi_d = h_d(\alpha)$
Linear predictor	v	$G_m\gamma_m$	b	$G_d\gamma_d$
Deviance	d_m	gamma(d_m^*, λ)	d_d	gamma(d_d^*, α)
Prior weight	$1/\lambda$	$(1-q_m)/2$	$1/\alpha$	$(1-q_d)/2$

†$d_i = 2\int_{\hat{y}_i}^{y} (y-s)/V(s)\,\mathrm{d}s$, $d_{mi} = 2\int_{\hat{u}_i}^{\psi} (\psi - s)/V_m(s)\,\mathrm{d}s$, $d_{di} = 2\int_{\hat{a}_i}^{\psi_d} (\psi_d - s)/V_d(s)\,\mathrm{d}s$, $d^* = d/(1-q)$, $d_m^* = d_m/(1-q_m)$, $d_d^* = d_d/(1-q_d)$, gamma$(d^*, \phi) = 2\{-\log(d^*/\phi) + (d^* - \phi)/\phi\}$ and (q, q_m, q_d) extends the idea of leverage to HGLMs (Lee and Nelder, 2001a).

Table 7. Components of GLM-type classes of model

Model	*Components*
GLM	β
JGLM†	β, γ
EGLM	β, u, λ
HGLM	$\beta, \gamma, u, \lambda$
DHGLM	$\beta, \gamma, u, \lambda, a, \alpha$

†The joint GLM of Section 5.2.

$$E(y) = \mu,$$
$$\mathrm{var}(y) = \phi\, V(\mu).$$

Nelder and Wedderburn (1972) introduced the GLM for the mean μ as follows:

$$\eta = g(\mu)$$
$$= X\beta.$$

A GLM is a special case of a DHGLM (Table 6) with a β-component only, assuming a one-parameter exponential family, whose log-likelihood is proportional to

$$\sum\{y\theta - k(\theta)\}/\phi, \tag{9}$$

where $\theta = \theta(\mu)$ is the canonical parameter and $k(\cdot)$ is the cumulant-generating function of y. The ML estimators for β can be obtained from the IWLS equations

$$X^{\mathrm{T}}\Sigma^{-1}X\beta = X^{\mathrm{T}}\Sigma^{-1}z, \tag{10}$$

where $z = \eta + (y-\mu)(\partial\eta/\partial\mu)$ is the adjusted dependent variable and $\Sigma = \Phi W^{-1}$ with

$$W = (\partial\mu/\partial\eta)^2\, V(\mu)^{-1}$$

and $\Phi = \phi I$. A GLM is specified by a response variable y, a variance function $V(\cdot)$, a link function $g(\cdot)$, a linear predictor $X\beta$ and a prior weight $1/\phi$. Within GLMs the variance function is sufficient to characterize a family of distributions for the response variable.

5.2. Joint generalized linear models and interconnected generalized linear models for mean and dispersion

In GLMs ϕ is assumed to be constant. Nelder and Lee (1991) introduced joint GLMs of independent observations by allowing another GLM for ϕ:

$$\begin{aligned}\xi &= h(\phi)\\ &= G\gamma.\end{aligned} \tag{11}$$

This formulation is especially relevant in quality improvement experiments where the aim is, for example, to minimize variance while holding the mean at a fixed target value.

A joint GLM is a special case of a DHGLM with β- and γ-components only, so that $v = 0$ and $b = 0$. If y has a normal distribution, this model can be fitted conveniently via two interconnected GLMs for the mean and dispersion as follows.

(a) The mean GLM: given $\hat{\phi}$, we estimate β by the IWLS equations (10) for the GLM, characterized by a response y, a normal error, an identity link, a linear predictor $X\beta$ and a prior weight $1/\hat{\phi}$.

(b) The dispersion GLM: given $\hat{\beta}$, we estimate γ by the IWLS equations

$$G^{\mathrm{T}}\Sigma_d^{-1}G\gamma = G^{\mathrm{T}}\Sigma_d^{-1}z_d, \tag{12}$$

where $\Sigma_d = \Gamma_d W_d^{-1}$ with $\Gamma_d = \mathrm{diag}\{2/(1-q_i)\}$, q_i is the ith element of the GLM leverages, i.e. the ith diagonal element of $X(X^{\mathrm{T}}\Sigma^{-1}X)^{-1}X^{\mathrm{T}}\Sigma^{-1}$, $\Sigma = \Phi W^{-1}$, $W = (\partial\mu/\partial\eta)^2\, V(\mu)^{-1}$, $\Phi = \mathrm{diag}(\phi_i)$, the weight functions $W_d = \mathrm{diag}(W_{di})$ with

$$W_{di} = \frac{(\partial\phi_i/\partial\xi_i)^2}{2\phi_i^2},$$

the dependent variables z_d with $z_{di} = \xi_i + (d_i^* - \phi_i)(\partial\xi_i/\partial\phi_i)$, $d_i^* = d_i/(1-q_i)$ and

$$d_i = 2\int_{\hat{\mu}_i}^{y} (y-s)/V(s)\,\mathrm{d}s = (y_i - \hat{\mu}_i)^2$$

is the GLM deviance component. This GLM is characterized by a response d^*, a gamma error, a link $h(\cdot)$, a linear predictor G and a prior weight $(1-q)/2$.

This method gives the ML estimators for β and the REML estimators for ϕ. For the link function $h(\cdot)$ we usually take the logarithm, which ensures that the dispersion components will be positive. Other links with this property could be used, but there is rarely enough information in the data to distinguish them from the log-link. For other GLM families of models, allowing

an exact likelihood, (extended) REML estimators can be obtained from equation (12), by modifying the GLM leverage q in the algorithm: for an illustration see Section 6.1 for gamma GLMs. For general variance functions that do not allow an exact likelihood, we can use, for estimating γ, either the extended quasi-likelihood (EQL) of Nelder and Pregibon (1987) or the pseudolikelihood of Davidian and Carroll (1988), based on a normal likelihood. For normal errors with $V(\mu)=1$, they are the same, but in general the EQL estimator from equation (12) is different from the pseudolikelihood estimator, which can be obtained by simply using the Pearson deviance $d_i=(y_i-\hat{\mu}_i)^2/V(\hat{\mu}_i)$ in equation (12) (Lee and Nelder, 1998). Pseudolikelihood estimators are consistent (Davidian and Carroll, 1988), but their asymptotic consistency may often be offset by the large mean-square error in finite samples (Nelder and Lee, 1992).

5.3. Hierarchical generalized linear models and augmented generalized linear models

An HGLM is a DHGLM with $b=0$ (i.e. without the a- and α-components) and assumes that $E(y|u)=\mu$ and $\mathrm{var}(y|u)=\phi\, V(\mu)$. In conjugate HGLMs the log-likelihood kernel for $v=\theta(u)$ becomes (Lee and Nelder, 2001a)

$$\sum\{\psi v-k(v)\}/\lambda,$$

so that from equation (9) ψ (known constants for each conjugate distribution) can be treated as quasi-responses and u (and hence $v=\theta(u)$) as quasi-fixed parameters, satisfying the purely formal relationships

$$E(\psi)=u,$$
$$\mathrm{var}(\psi)=\lambda\, V(u).$$

In addition to conjugate models, various combinations of GLM distributions and links for $y|v$ and any conjugate GLM distribution and link for v can be used to construct HGLMs. An HGLM can be viewed as an *augmented GLM* with the response variables $(y^{\mathrm{T}},\psi_m^{\mathrm{T}})^{\mathrm{T}}$, assuming

$$\mu=E(y),$$
$$u=E(\psi_m),$$
$$\mathrm{var}(y)=\phi\, V(\mu),$$
$$\mathrm{var}(\psi_m)=\lambda\, V_m(u)$$

and the linear predictor

$$\eta_{ma}=(\eta^{\mathrm{T}},\eta_m^{\mathrm{T}})^{\mathrm{T}}=T_m\omega,$$

where $\eta=g(\mu)=X\beta+Zv$, $\eta_m=g_m(u)=v$, $\omega=(\beta^{\mathrm{T}},v^{\mathrm{T}})^{\mathrm{T}}$ and

$$T_m=\begin{pmatrix} X & Z \\ 0 & I \end{pmatrix}.$$

For example, a GLMM is the HGLM with a constant variance function $V_m(u)=1$. Thus ω can be estimated by IWLS equations

$$T_m^{\mathrm{T}}\Sigma_{ma}^{-1}T_m\omega=T_m^{\mathrm{T}}\Sigma_{ma}^{-1}z_{ma}, \tag{13}$$

where $z_{ma}=(z^{\mathrm{T}},z_m^{\mathrm{T}})^{\mathrm{T}}$ and $\Sigma_{ma}=\Gamma_{ma}W_{ma}^{-1}$ with $\Gamma_{ma}=\mathrm{diag}(\Phi,\Lambda)$, $\Phi=\mathrm{diag}(\phi_i)$ and $\Lambda=\mathrm{diag}(\lambda_i)$. The dependent variables $z_{mai}=(z_i,z_{mi})$ are defined by

$$z_i=\eta_i+(y_i-\mu_i)(\partial\eta_i/\partial\mu_i)$$

for the data y_i, and

$$z_{mi} = v_i + (\psi_m - u_i)(\partial v_i / \partial u_i)$$

for the quasi-response ψ_m. The weight functions $W_{ma} = \text{diag}(W_{m0}, W_{m1})$ are defined by

$$W_{m0i} = (\partial \mu_i / \partial \eta_i)^2 \, V(\mu_i)^{-1}$$

for the data y_i and

$$W_{m1i} = (\partial u_i / \partial v_i)^2 \, V_m(u_i)^{-1}$$

for the quasi-response ψ_m. Equation (13) generalizes Henderson's (1975) mixed model equation.

In Table 6, the component GLM for λ is a dispersion GLM for the u-component. Thus, fitting of two interconnected GLMs for joint GLMs can be applied to estimate γ_m, giving the IWLS equations

$$G_m^{\mathrm{T}} \Sigma_m^{-1} G_m \gamma_m = G_m^{\mathrm{T}} \Sigma_m^{-1} z_m, \tag{14}$$

where $\Sigma_m = \Gamma_m W_m^{-1}$ with $\Gamma_m = \text{diag}\{2/(1 - q_{mi})\}$, $W_m = \text{diag}(W_{mi})$ with

$$W_{mi} = \frac{(\partial \lambda_i / \partial \xi_{mi})^2}{2\lambda_i^2}$$

and z_m with

$$z_{mi} = \xi_{mi} + (d_{mi}^* - \lambda_i)(\partial \xi_{mi} / \partial \lambda_i)$$

where

$$d_{mi}^* = d_{mi} / (1 - q_{mi}),$$

$$d_{mi} = 2 \int_{\hat{u}_i}^{\psi} (\psi - s) / V_m(s) \, \mathrm{d}s;$$

q_{mi} extends the idea of leverage to HGLMs (Lee and Nelder, 2001a). This GLM is characterized by a response d_m^*, a gamma error, a link $h_m(\cdot)$, a linear predictor G_m and a prior weight $(1 - q_m)/2$.

It is the use of the h-likelihood for estimation that leads to fitting of interlinked IWLS equations for fixed and random parameters.

5.4. Correlated random effects and linear transformations

Independent random components allow only positive correlations. We can have arbitrary covariance or precision matrices by considering a linear transformation of the random effects (Lee and Nelder, 2001b) $v = L(\rho) r$ where the elements r are independent, with joint distribution given by

$$r = L(\rho)^{-1} v \sim \text{MVN}(0, \Lambda)$$

with $\Lambda = \text{diag}(\lambda_i)$; MVN stands for the multivariate normal distribution. Here ρ denotes parameters for temporal and spatial correlations and can be estimated by using an HGLM

$$\begin{aligned} \eta &= g(\mu) \\ &= X\beta + Z^* r, \end{aligned}$$

where $Z^* = Z\,L(\rho)$, and in which Z^* is updated iteratively. It is not necessary to impose a Gaussian assumption on r, but it gives a meaningful interpretation as correlations. The state

space models of Harvey (1989) and Durbin and Koopman (2000) and the spatial models of Besag and Higdon (1999) are models in which $L(\rho) = L$, i.e. there are no unknown parameters. In forming $L(\rho)$ the use of regression models was studied by Pourahmadi (2000) and Pan and MacKenzie (2003).

5.5. Fitting double hierarchical generalized linear models

Similarly to HGLMs, the distribution of random effects a for the dispersion model can be characterized by the variance function $V_d(a)$. In this paper a conjugate DHGLM is a model having

$$V_m(u) = V(u),$$
$$V_d(a) = a^2.$$

The inverse gamma distribution appears as the conjugate of the gamma distribution, which is the distribution of the responses d^* for the dispersion model.

To obtain an insight into the fitting algorithm for DHGLMs in Section 2 we first consider the SV model (8), a simple DHGLM with $\mu = 0$. Here the h-likelihood becomes

$$\begin{aligned} h &= \log\{f(y|b;\phi)\} + \log\{f(b;\alpha)\} \\ &= -\tfrac{1}{2}[\textstyle\sum\{d_t/\phi_t + \log(2\pi\phi_t)\} + \sum\{r_i^2/\alpha + \log(2\pi\alpha)\}] \end{aligned}$$

where $d_t = y_t^2$, $\log(\phi_t) = \gamma + b_t = \gamma + L_t r$, X_t and L_t are tth rows of the model matrices X and L ($= L(\rho)$ in equation (4)) and r is the vector of r_i.

As estimation criteria we use h for (r, γ) and $p_\gamma(L_b)$ for (γ_d, ρ). Because the use of L_b often involves intractable integration we propose to use $p_{b,\gamma}(h)$, as shown in Table 8. This strategy gives the following estimating procedure.

(a) For estimating $\psi = (\gamma, r)$ the IWLS equations (12) are extended to those for an augmented GLM

$$T_d^{\mathrm{T}} \Sigma_{da}^{-1} T_d \psi = T_d^{\mathrm{T}} \Sigma_{da}^{-1} z_{da}, \tag{15}$$

where

$$T_d = \begin{pmatrix} G & L \\ 0 & I_q \end{pmatrix},$$

$\Sigma_{da} = \Gamma_{da} W_{da}^{-1}$ with $\Gamma_{da} = \mathrm{diag}\{2/(1-q), \Psi\}$, $\Psi = \mathrm{diag}(\alpha_i)$ and the weight functions $W_{da} = \mathrm{diag}(W_{d0}, W_{d1})$ with

$$W_{d0i} = \frac{(\partial\phi_i/\partial\xi_i)^2}{2\phi_i^2}$$

Table 8. Criteria for effects in SV models

Criterion	*Arguments*	*Estimated*	*Eliminated*	*Approximation*
h	$r, \gamma, \gamma_d, \rho$	r, γ	None	h
$p_\gamma(L_b)$	γ_d, ρ	γ_d, ρ	b, γ	$p_{b,\gamma}(h)$

for the gamma data d_i^*, and $W_{d1i} = (\partial a_i/\partial r_i)^2\, V_d(a_i)^{-1}$ for the quasi-response ψ_d, and the dependent variables $z_{da} = (z_{d0}, z_{d1})$ with

$$z_{d0i} = \xi_i + (d_i^* - \phi_i)(\partial \xi_i/\partial \phi_i)$$

for the data d_i^* and

$$z_{d1i} = r_i + (\psi_d - r_i)(\partial r_i/\partial a_i)$$

for the quasi-response ψ_d. In SV models, $q = 0$ because there is no $(\beta^{\mathrm{T}}, v^{\mathrm{T}})^{\mathrm{T}}$,

$$z_{d0i} = \xi_i + (d_i^* - \phi_i)/\phi_i$$

for the log-link $\xi_i = \log(\phi_{ii}^*)$ and

$$z_{d1i} = r_i + (\psi_d - r_i)(\partial r_i/\partial a_i) = 0$$

for $\psi_d = 0$ and $r_i = a_i$.

(b) For estimating γ_d we have IWLS estimation similar to equations (14)

$$G_d^{\mathrm{T}} \Sigma_{d1}^{-1} G_d \gamma_d = G_d^{\mathrm{T}} (I - Q_d) \Sigma_{d1}^{-1} z_d, \tag{16}$$

where $Q_d = \mathrm{diag}(q_{di})$, $\Sigma_{d1} = \Gamma_d W_d^{-1}$, $\Gamma_d = \mathrm{diag}\{2/(1 - q_d)\}$, the weight functions $W_d = \mathrm{diag}(W_{di})$ with

$$W_{di} = \frac{(\partial \alpha_i/\partial \xi_{di})^2}{2\alpha_i^2},$$

the dependent variables z_d with

$$z_{di} = \xi_{di} + (d_{di}^* - \alpha_i)(\partial \xi_{di}/\partial \alpha_i)$$

and the deviance components

$$d_{di} = 2 \int_{a_i}^{\psi_d} (\psi_d - s)/V_d(s)\, \mathrm{d}s;$$

q_d extends the idea of leverage to HGLMs (Lee and Nelder, 2001a). For estimating ρ we use Lee and Nelder's (2001b) method in Section 5.4. This algorithm is equivalent to that for gamma HGLMs with responses y_t^2.

As estimation criteria for DHGLMs we use h for (v, β), $p_\beta(L_v)$ for (b, γ, γ_m) and $p_{\beta,\gamma}(L_{b,v})$ for γ_d. Because the formation of L_v and $L_{b,v}$ often involves intractable integration, we propose to use $p_{v,\beta}(h)$ and $p_{v,\beta,b,\gamma}(h)$, instead of $p_\beta(L_v)$ and $p_{\beta,\gamma}(L_{b,v})$, as shown in Table 9. This strategy gives the following estimating procedure.

(a) For estimating $\omega = (\beta^{\mathrm{T}}, v^{\mathrm{T}})^{\mathrm{T}}$, use the IWLS equations (13).
(b) For estimating γ_m, use the IWLS equations (14).
(c) For estimating $\psi = (\gamma^{\mathrm{T}}, b^{\mathrm{T}})^{\mathrm{T}}$, use the IWLS equations (15), with $L = 0$ for T_d.
(d) For estimating γ_d, use the IWLS equations (16).

We can accommodate temporal and spatial random effects for both equations (12) and equations (15) by using L-matrices as in Section 5.4. This completes the explanation of the fitting algorithm for DHGLMs in Table 6.

For binary data the use of $p_v(h)$ for estimation of β can be accommodated by modifying the adjusted dependent variables z_{ma} in equation (13) (Noh and Lee, 2004) and the use of

Table 9. Criteria for effects in DHGLMs

Criterion	*Arguments*	*Estimated*	*Eliminated*	*Approximation*
h	$v, \beta, b, \gamma, \gamma_m, \gamma_d$	v, β	None	h
$p_\beta(L_v)$	$b, \gamma, \gamma_m, \gamma_d$	b, γ, γ_m	v, β	$p_{v,\beta}(h)$
$p_{\beta,\gamma}(L_{b,v})$	γ_d	γ_d	v, β, b, γ	$p_{v,\beta,b,\gamma}(h)$

second-order approximations can easily be accommodated by modifying the leverage q_m in equations (14) (Lee and Nelder, 2001a).

6. Extensions

We have seen how the three fitting methods can be combined to fit DHGLMs. Further extensions of the fitting algorithm are possible. For example the IWLS procedures are easily extended to multivariate DHGLMs as shown in Lee *et al.* (2005).

6.1. Quasi-double hierarchical generalized linear models and extended quasi-likelihood

In HGLMs, the GLMs both for the $y|v$ and for the u-components can be extended to quasi-likelihood models, which are characterized solely by variance functions $V(\mu)$ and $V_m(u)$, for which GLM families of distributions need not exist. For inferences from such quasi-HGLMs Lee and Nelder (2001a) proposed to use the double EQL q, which approximate the likelihoods for the $y|v$ and v by EQLs. Given $(\hat{\beta}(\tau), \hat{v}(\tau))$, the maximization of $p_{\beta,v}(q)$ for $\tau = (\gamma, \gamma_m)$ leads to the gamma GLM equations (12) and (14), but with (q, q_m) being replaced by the GLM leverage q^* from the augmented GLM for the mean (13); q_i^* is the ith GLM leverage, i.e. the ith diagonal element of

$$T_m^{\mathrm{T}}(T_m^{\mathrm{T}}\Sigma_m^{-1}T_m)^{-1}T_m^{\mathrm{T}}\Sigma_m^{-1}.$$

An advantage of using EQL is that a single code can be used for a broader class of models. Similarly our method can be applied to quasi-DHGLMs, which are characterized by $(V(\mu), V_m(u), V_d(a))$. In finite samples double EQL estimators often have smaller mean-square errors than ML estimators: see Lee (2004) for binary data and Saha and Paul (2005) for count data.

Since the EQL is not a true likelihood, it may not give consistent estimators. This can be avoided for models that allow true likelihoods for all the component GLMs. For example, suppose that the $y|v$ component follows the gamma GLM such that $E(y|v) = \mu$ and $\mathrm{var}(y|v) = \phi\mu^2$; we have

$$-2\log\{f(y|v, b; \beta, \phi)\} = \sum[d_i/\phi_i + 2/\phi_i + 2\log(\phi_i)/\phi_i + 2\log\{\Gamma(1/\phi_i)\}],$$

where

$$d_i = 2\int_{\mu_i}^{y_i} (y_i - s)/s^2 \, \mathrm{d}s = 2\{(y-\mu)/\mu - \log(y)/\mu\}.$$

The corresponding EQL is

$$-2\log\{q(y|v, b; \beta, \phi)\} = \sum\{d_i/\phi_i + \log(2\pi\phi_i y_i^2)\}.$$

Now $\log\{f(y|v,b;\beta,\phi)\}$ and $\log\{q(y|v,b;\beta,\phi)\}$ are equivalent up to the Stirling approximation

$$\log\{\Gamma(1/\phi_i)\} \approx -\log(\phi_i)/\phi_i + \log(\phi_i)/2 + \log(2\pi)/2 - 1/\phi_i.$$

Thus, the EQL gives bias when the value of ϕ is large (for which the above approximation is bad). It can be shown that $\partial p_{\beta,v}(h)/\partial\gamma_k = 0$ leads to the IWLS equations (12) with

$$q_i = q_i^* + 1 + \frac{2\log(\phi_i)}{\phi_i} + \frac{2\,\mathrm{dg}(1/\phi_i)}{\phi_i},$$

where $\mathrm{dg}(\cdot)$ is the digamma function. We can find q_{mi} for equations (14) similarly. This shows how to use the GLM attributes in Table 6 to develop an algorithm for fitting models by using h-likelihood. Thus we use q_i as an extended leverage for gamma GLM fitting.

6.2. Heavy-tailed distributions for random effects

Wakefield *et al.* (1994) proposed, in a Bayesian setting, the use of a multivariate t-distribution for random effects in μ and found that the resulting model also gives robust estimation against outliers. This is equivalent to introducing random effects in the linear predictor (2) of the λ-component. There has been concern about the choice of distribution, because of the difficulty in identification from limited data, especially binary data. The nonparametric ML estimator can be fitted by assuming discrete latent distributions (Laird, 1978) and its use was recommended by Heckman and Singer (1984) because the parameter estimates in random-effect models can be sensitive to misspecification; see also Schumacher *et al.* (1987). However, its use has been restricted by the difficulties in fitting discrete latent models, e.g. in choosing the number of discrete points (McLachlan, 1987) and in computing standard errors for nonparametric ML (McLachlan and Krishnan, 1997; Aitkin and Alfo, 1998). Noh *et al.* (2004) showed that the introduction of random effects in the λ-component removes such sensitivity in parameter estimation to the choice of distribution of random effects.

6.3. Multivariate double hierarchical generalized linear models and other extensions

The h-likelihood approach has been extended to joint DHGLMs for multivariate responses. Ha *et al.* (2003) considered the use of joint modelling of repeated measures and survival times, and Yun and Lee (2004a) that for continuous and binary responses. This covers the missing data problem by taking the missingness indicator as a binary response: see Lee *et al.* (2005) for shared random-effect models and Yun and Lee (2004b) for selection models. Ma *et al.* (2003) and Ha and Lee (2005) showed that nonparametric base-line frailty models can be fitted by using Poisson HGLMs. Thus, structured dispersion and heavy tails can be introduced into frailty models. Noh *et al.* (2004) and Yun and Lee (2004b) showed numerically that the h-likelihood estimation method provided statistically efficient estimation for the DHGLMs that they considered.

In non-linear mixed effects models Lindstrom and Bates (1990), Wolfinger (1993) and Vonesh (1996) have developed IWLS equations that are similar to ours. Recently, Vonesh *et al.* (2002) have applied the h-likelihood approach in Section 5.1 to non-linear mixed effects models with correlated Gaussian errors and correlated Gaussian random effects. Such models are often encountered in pharmacokinetic studies. These algorithms can be similarly extended to allow heavy tails.

7. Concluding remarks

DHGLMs cover a broad spectrum of distributions and provide great flexibility in dispersion modelling. With the h-likelihood apparatus such extensive classes of new models can be brought

together within a single framework. Because the h-likelihood method is numerically efficient it is a practical choice for computing the ML and REML estimators for an extended class of models. Yun and Lee (2004b) found that, for a DHGLM with a single random component each for the mean and dispersion, the h-likelihood procedure took less than 8 min on a personal computer with a Pentium 4 processor and 526 Mbytes of random access memory. Because there are only two random components, numerical integration such as Gauss–Hermite quadrature could be used. With the SAS `NLMIXED` procedure using adaptive Gauss–Hermite quadrature with 20 and 25 quadrature points the time exceeded 35 and 57 h respectively. With 40 and 50 quadrature points the algorithm failed after 3 and 5 days respectively.

We have proposed to use h-likelihood for inference from general models having unobserved or unobservable random variables. h-likelihood has provided statistically efficient estimates for all the models where we have done numerical studies. However, we are not claiming the axiomatic use of h-likelihood for all extended models. Its behaviour should be thoroughly investigated model by model. If some version of the h-likelihood method is not working satisfactorily in some model it may be possible to modify it to reduce the bias etc., e.g. by using a higher order approximation to reduce the bias or total bias elimination. However, we have found no method (including MCMC-type methods) which is better than the h-likelihood procedure in bias even with extreme binary data if a proper version is used (Noh and Lee, 2004). For example MCMC sampling is difficult to apply to large data sets (Noh *et al*., 2004) and it is biased in small samples (Noh and Lee, 2004). The Gauss–Hermite quadrature method cannot be applied to crossed models, whereas the h-likelihood method can. Thus, the h-likelihood method is worth attention. It does not require the possibly difficult E-step in the EM method to obtain the ML estimators.

For DHGLMs we have developed Genstat code to allow arbitrary combinations of GLM distributions and links for $y|v$ and any conjugate GLM distributions and links for v and b. Random effects v and b may have crossed or nested structures and also temporal and spatial correlations. Software can be obtained from `j.nelder@imperial.ac.uk`.

Acknowledgement

This research was supported by a grant from the Statistical Research Center for Complex Systems of the Korean Science and Engineering Foundation.

References

Aitkin, M. and Alfo, M. (1987) Regression models for binary longitudinal responses. *Statist. Comput.*, **8**, 289–307.
Besag, J. and Higdon, P. (1999) Bayesian analysis of agricultural field experiments (with discussion). *J. R. Statist. Soc.* B, **61**, 691–746.
Bjørnstad, J. F. (1996) On the generalization of the likelihood function and likelihood principle. *J. Am. Statist. Ass.*, **91**, 791–806.
Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear mixed models. *J. Am. Statist. Ass.*, **88**, 9–25.
Breslow, N. E. and Lin, X. (1995) Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, **82**, 81–91.
Brumback, B. A. and Rice, J. A. (1998) Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *J. Am. Statist. Ass.*, **93**, 961–994.
Chernoff, H. (1954) On the distribution of the likelihood ratio. *Ann. Math. Statist.*, **25**, 573–578.
Cox, D. R. and Reid, N. (1987) Parameter orthogonality and approximate conditional inference. *J. R. Statist. Soc.* B, **49**, 1–18.
Davidian, M. and Carroll, R. J. (1988) A note on extended quasi-likelihood. *J. R. Statist. Soc.* B, **50**, 74–82.
Diggle, P. J., Liang, K. and Zeger, S. L. (1994) *Analysis of Longitudinal Data*. Oxford: Clarendon.
Durbin, J. and Koopman, S. J. (2000) Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives (with discussion). *J. R. Statist. Soc.* B, **62**, 3–56.
Engel, R. E. (1995) *ARCH*. Oxford: Oxford University Press.

Goldstein, H. (1995) *Multilevel Statistical Models*. London: Arnold.
Green, P. J. and Silverman, B. W. (1994) *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach*. London: Chapman and Hall.
Ha, I. D. and Lee, Y. (2005) Comparison of hierarchical likelihood versus orthodox BLUP approach for frailty models. *Biometrika*, **92**, 717–723.
Ha, I. D., Lee, Y. and Song, J. K. (2001) Hierarchical likelihood approach for frailty models. *Biometrika*, **88**, 233–243.
Ha, I. D., Lee, Y. and Song, J. K. (2002) Hierarchical likelihood approach for mixed linear models with censored data. *Liftime Data Anal.*, **8**, 163–176.
Ha, I. D., Park, T. and Lee, Y. (2003) Joint modelling of repeated measures and survival time data. *Biometr. J.*, **45**, 647–658.
Harvey, A. C. (1989) *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.
Harvey, A. C., Ruiz, E. and Shephard, N. (1994) Multivariate stochastic variance models. *Rev. Econ. Stud.*, **61**, 247–264.
Harville, D. (1977) Maximum likelihood approaches to variance component estimation and related problems. *J. Am. Statist. Ass.*, **72**, 320–340.
Heckman, J. and Singer, B. (1984) A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, **52**, 271–320.
Henderson, C. R. (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics*, **31**, 423–447.
Hougaard, P. (2000) *Analysis of Multivariate Survival Data*. New York: Springer.
Hudak, S. J., Saxena, A., Bucci, R. J. and Malcom, R. C. (1978) Development of standard methods of testing and analyzing fatigue crack growth rate data. *Technical Report AFML-TR-78-40*. Westinghouse R&D Center, Westinghouse Electric Corporation, Pittsburgh.
Kim, S., Shephard, N. and Chib, S. (1998) Stochastic volatility: likelihood inference and comparison with ARCH models. *Rev. Econ. Stud.*, **98**, 361–393.
Laird, N. (1978) Nonparametric maximum likelihood estimation of a mixing distribution. *J. Am. Statist. Ass.*, **73**, 805–811.
Lange, K. L., Little, R. J. A. and Taylor, J. M. G. (1989) Robust statistical modeling using the t Distribution. *J. Am. Statist. Ass.*, **84**, 881–896.
Lawless, J. and Crowder, M. (2004) Covariates and random effects in a gamma process model with application to degeneration and failure. *Liftime Data Anal.*, **10**, 213–227.
Lee, Y. (2001) Can we recover information from concordant pairs in binary matched paired? *J. Appl. Statist.*, **28**, 239–246.
Lee, Y. (2004) Estimating intraclass correlation for binary data using extended quasi-likelihood. *Statist. Modllng*, **4**, 113–126.
Lee, Y. and Nelder, J. A. (1996) Hierarchical generalized linear models (with discussion). *J. R. Statist. Soc.* B, **58**, 619–678.
Lee, Y. and Nelder, J. A. (1998) Generalized linear models for the analysis of quality-improvement experiments. *Can. J. Statist.*, **26**, 95–105.
Lee, Y. and Nelder, J. A. (2000) Two ways of modelling overdispersion in non-normal data. *Appl. Statist.*, **49**, 591–598.
Lee, Y. and Nelder, J. A. (2001a) Hierarchical generalised linear models: a synthesis of generalised linear models, random effect models and structured dispersions. *Biometrika.*, **88**, 987–1006.
Lee, Y. and Nelder, J. A. (2001b) Modelling and analysing correlated non-normal data. *Statist. Modllng*, **1**, 3–16.
Lee, Y. and Nelder, J. A. (2005a) Fitting via alternative random-effect models. *Statist. Comput.*, to be published.
Lee, Y. and Nelder, J. A. (2005b) Likelihood for random-effect models (with discussion). *Statist. Oper. Res. Trans.*, **29**, 141–182.
Lee, Y., Noh, M. and Ryu, K. (2005) HGLM modeling of dropout process using a frailty model. *Comput. Statist.*, **20**, 295–309.
Lin, X. and Breslow, N. E. (1996) Bias correction in generalised linear mixed models with multiple components of dispersion. *J. Am. Statist. Ass.*, **91**, 1007–1016.
Lindsey, J. K. and Lindsey, P. J. (2006) Multivariate distributions with correlation matrices for repeated measurements. *Comput. Statist. Data. Anal.*, **50**, 720–732.
Lindstrom, M. J. and Bates, D. B. (1990) Nonlinear mixed effects models for repeated measures data. *Biometrics*, **46**, 673–687.
Longford, N. (1993) *Random Coefficient Models*. Oxford: Oxford University Press.
Lu, C. J. and Meeker, W. Q. (1993) Using degeneration measurements to estimate a time-to-failure distribution. *Technometrics*, **35**, 161–174.
Ma, R., Krewski, D. and Burnett, R. T. (2003) Random effects Cox models: a Poisson modelling approach. *Biometrika*, **90**, 157–169.
McLachlan, G. J. (1987) On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Appl. Statist.*, **36**, 318–324.

McLachlan, G. J. and Krishnan, T. (1997) *The EM Algorithm and Extensions*. New York: Wiley.
Medina, A., Lakhina, H., Matta, I. and Byers, J. (2001) BRITE: universal topology generation from a user's perspective. *Technical Report*. Boston University, Boston.
Nelder, J. A. and Lee, Y. (1991) Generalized linear models for the analysis of Taguchi-type experiments. *Appl. Stochast. Mod. Data Anal.*, **7**, 107–120.
Nelder, J. A. and Lee, Y. (1992) Likelihood, quasi-likelihood and pseudolikelihood: some comparisons. *J. R. Statist. Soc.* B, **54**, 273–284.
Nelder, J. A. and Pregibon, D. (1987) An extended quasi-likelihood function. *Biometrika*, **74**, 221–231.
Nelder, J. A. and Wedderburn, R. W. M. (1972) Generalized linear models. *J. R. Statist. Soc.* A, **135**, 370–384.
Noh, M. and Lee, Y. (2004) REML estimation for binary data in generalized linear mixed models. To be published.
Noh, M., Lee, Y. and Pawitan, Y. (2005) Robust ascertainment-adjusted parameter estimation. *Genet. Epidem.*, **29**, 68–75.
Noh, M., Yip, Y., Lee, Y. and Pawitan, Y. (2004) Multicomponent variance estimation for binary traits in family-based studies. *Genet. Epidem.*, to be published.
Pan, J. X. and MacKenzie, G. (2003) On modelling mean-covariance structures in longitudinal studies. *Biometrika*, **90**, 239–244.
Patil, G. P. (1963) A characterization of the exponential-type distribution. *Biometrika*, **50**, 205–207.
Pawitan, Y. (2001) *In All Likelihood: Statistical Modelling and Inference using Likelihood*. Oxford: Oxford University Press.
Pourahmadi, M. (2000) Maximum likelihood estimation of generalized linear models for multivariate normal covariance matrix. *Biometrika*, **87**, 425–435.
Rigby, R. A. and Stasinopoulos, D. M. (1996) MADAM macros to fit mean and dispersion additive models. In *GLIM4 Macro Library Manual, Release 2.0* (eds A. Scallan and G. Morgan), pp. 48–84. Oxford: Numerical Algorithms Group.
Robinson, G. K. (1991) That BLUP is a good thing: the estimation of random effects. *Statist. Sci.*, **6**, 15–51.
Robinson, M. E. and Crowder, M. (2000) Bayesian methods for a growth-curve degradation model with repeated measures. *Liftime Data Anal.*, **6**, 357–374.
Saha, K. and Paul, S. (2005) Bias corrected maximum likelihood estimator of negative binomial dispersion parameter. *Biometrics*, **61**, 179–185.
Schall, R. (1991) Estimation in generalized linear models with random effects. *Biometrika*, **40**, 917–927.
Schumacher, M., Olschewski, M. and Schmoor, C. (1987) The impact of heterogeneity on the comparison of survival times. *Statist. Med.*, **6**, 773–784.
Shephard, N. (1996) Statistical aspects of ARCH and stochastic volatility. In *Time Series Models in Econometrics, Finance and Other Fields* (eds D. R. Cox, O. E. Barndorff-Nielsen and D. V. Hinkley). London: Chapman and Hall.
Shephard, N. and Pitt, M. R. (1997) Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, **84**, 653–667.
Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *J. R. Statist. Soc.* B, **64**, 583–639.
Thall, P. F. and Vail, S. C. (1990) Some covariance models for longitudinal count data with overdispersion. *Biometrics*, **46**, 657–671.
Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data*. Berlin: Springer.
Verbyla, A. P., Cullis, B. R., Kenward, M. G. and Welham, S. J. (1999) The analysis of designed experiments and longitudinal data by using smoothing splines (with discussion). *Appl. Statist.*, **48**, 269–311.
Vonesh, E. F. (1996) A note on the use of Laplace's approximation for nonlinear mixed-effects models. *Biometrika*, **83**, 447–452.
Vonesh, E. F., Wang, H., Nie, L. and Majumdar, D. (2002) Conditional second-order generalized estimating equations for generalized linear and nonlinear mixed-effects models. *J. Am. Statist. Ass.*, **97**, 271–283.
Wahba, G. (1990) *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.
Wakefield, J. C., Smith, A. F. M., Racine-Poon, A. and Gelfand, A. E. (1994) Bayesian analysis of linear and non-linear population models by using the Gibbs sampler. *Appl. Statist.*, **43**, 201–221.
Wedderburn, R. W. M. (1974) Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika*, **61**, 439–447.
Welham, S., Cullis, B. R., Kenward, M. G. and Thompson, R. (2004) The analysis of longitudinal data using model L-splines. *Statist. Med.*, to be published.
Wolfinger, R. D. (1993) Covariance structure selection in general mixed models. *Communs Statist. Simuln Computn*, **22**, 1079–1106.
Yun, S. and Lee, Y. (2004a) Comparison of hierarchical and marginal likelihood estimators for binary outcomes. *Comput. Statist. Data. Anal.*, **45**, 639–650.
Yun, S. and Lee, Y. (2004b) Robust estimation in mixed linear models with non-monotone missingness. *Statist. Med.*, to be published.

Yun, S., Sohn, S. Y. and Lee, Y. (2004) Modeling non-homogeneous LRD queueing system with covariates: inverse gamma mixture of Pareto. *J. Appl. Statist.*, to be published.
Zeger, S. L. and Diggle, P. J. (1994) Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics*, **50**, 689–699.
Zimmerman, D. and Nunez-Anton, V. (2001) Parametric modelling of growth curve data: an overview. *Test*, **10**, 1–73.

Discussion on the paper by Lee and Nelder

Gilbert MacKenzie (*Limerick University*)

This is another modelling *tour de force* from Lee and Nelder which builds on developments stemming from their first seminal paper on hierarchical generalized linear models (HGLMs) (Lee and Nelder, 1996). In particular, tonight's material is an important synthesis of two relatively recent papers (Lee and Nelder, 2000, 2001) leading to the formalization of double hierarchical generalized linear models (DHGLMs). The scope of the work is ambitious, seeking, *inter alia*, to bring a substantial portion of statistical modelling under a common inferential, computational and conceptual framework. Its ambition is not confined merely to conventional modelling schemes since, manifestly, DHGLMs extend these, elegantly, in several different ways.

Of general interest to the modelling community is the manner in which DHGLMs facilitate joint mean–covariance modelling (and selection), especially in relation to heavy-tailed distributions. Curiously, the importance of joint mean–covariance modelling is not well understood and, accordingly, I digress a little to explain in the context of longitudinal clinical trials—an area that is clearly covered by the authors' modelling schema.

Conventionally, in this context considerable emphasis is placed on estimation of the mean structure and less on the covariance structure, between repeated measurements on the same subject. Often, the covariance structure is thought to be a 'nuisance parameter' or at least not to be of primary 'scientific interest' and is selected from a limited menu of structures contained in software packages. Increasingly, such ideas are considered *passé* as, from an inferential standpoint, the problem is symmetrical in both parameters μ and Σ.

It was Rao (1965) who showed in the context of the general linear model that $\hat{\beta}$ is Σ invariant when

$$\Sigma = X\Gamma X' + Q\Theta Q'$$

where Γ of order $p \times p$ and Θ of order $(p-m) \times (p-m)$ are positive definitive and Q is a $p \times (p-m)$ matrix orthogonal to X, i.e. $Q'X = 0$ and we have assumed exactly m repeated measurements over time.

When Σ is outside this class we may expect that a suboptimal choice of Σ may influence μ and vice versa. If so, one remedy is to search the joint mean–covariance model space $\{\mathcal{M} \times \mathcal{C}\}$, to determine the optimal model with estimators $(\hat{\mu}, \hat{\Sigma})$. Pan and MacKenzie (2003) and MacKenzie and Pan (2006) provide efficient algorithms for this approach in certain (simple) classes of joint generalized linear models (JGLMs), extended generalized linear models (EGLMs) and HGLMs. However, the authors' contribution is potentially more powerful, providing a wider choice of link functions and modelling covariance structures in the latent space, via their L-matrices (Section 5.4). Moreover, the inclusion of random effects in the dispersion is a major advance: not only because they accommodate heavy-tailed distributions, but also because, finally, the modelling of mean and covariance structures is rendered symmetric. There is also no need to 'marginalize' over the random-effects distributions! This is more than a computational advantage. The importance of these ideas will not be lost on the pharmaceutical industry where, increasingly, small effects may have important benefits. Although DHGLMs facilitate joint model selection, they also pose many new problems in search optimization, thus opening new avenues of research.

Inference in h-likelihood is based on a penalized likelihood

$$p_\theta(l) = \left(l - \frac{1}{2} \log\left[\det\left\{ \frac{A(l,\theta)}{2\pi} \right\} \right] \right) \Bigg|_{\theta=\hat{\theta}}$$

where $A(l,\theta) = -\partial^2 l/\partial\theta^2$ and $\hat{\theta}$ solves $\partial l/\partial\theta = 0$. The object $p_\theta(l)$ may be viewed as a nuisance parameter elimination function. Thus, when the nuisance θ is a set of fixed effects we eliminate as if by conditioning and when the nuisance is a set of random effects we eliminate as if by integration, by means of a first-order Laplace approximation to the marginal likelihood. Accordingly, integration (i.e. Markov chain Monte Carlo sampling) is rendered unnecessary in a wide class of statistical models. In 1996, with exponential growth in computer power, this property was not perceived to be particularly advantageous. However,

today, with ever-increasing model complexity, the avoidance of integration is a very rich prize indeed in modern biostatistical, bioinformatical and genomical applications, especially in relation to model selection and criticism.

One important class of models that is referenced in the paper, but not dealt with at any length, is that of frailty survival models. Here, also, the impact of h-likelihood is being felt (Ha *et al.*, 2001). More recently, Ha and Lee (2005) have established the formal correspondence between the semiparametric proportional hazards frailty model, SPPH $\sim$ log-normal, and the Poisson HGLMs of Ma *et al.* (2003). Moreover, non-proportional-hazard models have also been implemented using h-likelihood methods. For example, the generalized time-dependent logistic model has been generalized to, *inter alia*, a GTDL $\sim$ gamma frailty model, Here, the h-likelihood estimators agree exactly with the (analytically tractable) marginal likelihood estimators, and even in multivariate extensions to recurrent events (MacKenzie *et al.*, 2006).

One simply cannot fail to remark on the rich tapestry of model classes that are introduced and discussed in this paper (GLMs, JGLMs, EGLMs, HGLMs and DHGLMs) and all computable within one elegant, conceptual, algorithmic framework—a remarkable result, which will repay closer study.

Finally, here I have tried to concentrate on the visionary aspects and compass of the h-likelihood approach to statistical modelling. The authors are to be congratulated on the scope of their project and I take great pleasure in proposing the vote of thanks.

David Firth (*University of Warwick, Coventry*)

I have been kindly reminded of the tradition that the seconder of the vote of thanks is supposed to be 'more critical'. I shall try to do my duty!

On reading the paper I was quickly put in the right frame of mind for the task at hand. The authors complain in Section 4 that their earlier paper read to the Society was unreasonably criticized by discussants, in that the usefulness of h-likelihood

> 'had not been fully appreciated because the worst result, which came from matched binary data, was held to characterize the entire HGLM estimation scheme'.

This refers to comments from at least four discussants who contested the claim made in Lee and Nelder (1996), section 5.3, that

> '... the procedures developed here may be reliable and useful as an approximate inference even in the worst situations'.

The criticisms all urged caution in the interpretation of that claim. B. Engel and A. Keen wrote

> 'How far this problem extends beyond binary and binomial data is not clear, but it shows that the authors are wrong in their assumption in Section 5.3 that MHL may work even in the worst situations'.

D. Clayton wrote

> 'These [asymptotic] arguments only apply when the sample size increases faster than the number of random effects—an assumption which will be unrealistic in most of the practical situations in which the use of random effect models would be considered. The authors further speculate that these methods continue to perform well in less ideal situations, but this is rather doubtful....'

J. Kuha wrote

> '... in situations like this, where there is little information on each random effect, the authors' comment on the reliability and usefulness of MHLEs should be treated with caution',

to which I myself added

> '... the binary matched pairs situation mentioned by earlier discussants is but an extreme case, and clearly caution is needed also in less extreme situations'.

Those cautionary remarks were perfectly reasonable, and necessary in view of the extent of the apparent claims being made for h-likelihood.

The further assertion that is made in tonight's paper that the 1996 criticisms of h-likelihood derived from 'misunderstanding of the details of its specification' is puzzling. It will be evident from the excerpts above that any such misunderstanding was common to many if not all readers! Some clue perhaps comes from the subsequent softening of strong statements that were made in the 1996 paper, such as

'It is important to note that our procedures make no use of marginal likelihood and so involve no integration'

(page 620) and

'As we have shown, we can treat the h-likelihood as if it were an orthodox likelihood for the parameters β and v...'

(page 649). The revised proposal described in tonight's paper appears to be to use the marginal likelihood where available, in preference to h-likelihood. This is welcome progress. In situations where the full marginal likelihood is difficult to compute, there may sometimes be a useful role also for methods based on 'composite' likelihoods (e.g. Lindsay (1988), Cox and Reid (2004) and Bellio and Varin (2005)).

My remarks here will focus on the methodological core of the paper. The 'h-likelihood method' for hierarchical generalized linear models proceeds in two stages, broadly as follows.

(a) Using a specified function $v = v(u)$ of the random effects u, define the h-likelihood

$$h = f(y|\beta, v)\, f(v|\alpha).$$

(b) Estimate the unknowns by maximizing three different criterion functions derived from h—one criterion for each of β, v and α. The recommended criterion functions (Table 5) have changed somewhat from Lee and Nelder (1996), and special criteria are now also suggested when y is binary.

I wish to make three general remarks in relation to this scheme.

(i) The choice of v is not unimportant, and is arbitrary. Lee and Nelder suggest to choose $v(u)$ to be the transformation which makes fixed and random effects additive, but that is an ambiguous prescription. A simple illustration comes from consideration of a stripped-down version of an example from Lee and Nelder (2005a). Suppose that y given u is exponentially distributed with rate u, and that u is exponentially distributed with known rate 1. Since there are no fixed effects, additivity is achieved by *any* choice of v: e.g. $1/E(y|u) = u = 0 + v_1(u)$ with $v_1(u) \equiv u$, or $\log\{E(y|u)\} = -\log(u) = 0 + v_2(u)$ with $v_2(u) \equiv -\log(u)$. The corresponding h-likelihoods are different and yield appreciably different maximum h-likelihood estimators $\hat{u}_1 = 1/(y+1)$ and $\hat{u}_2 = 2/(y+1)$. The objection might perhaps be raised that this example is *too* simple, since it has no fixed parameters at all, but the basic problem—that additivity is not a property of the model but of the way that we choose to *write* the model—is a more general one. For example, if our model is $\{y_{ij}|u_i \sim \text{Poisson}(\beta u_i), u_i - 1 \sim \text{exponential}(\alpha)\}$ it can be alternatively written as $\{y_{ij}|u_i^* \sim \text{Poisson}(\beta + u_i^*), u_i^* \sim \text{exponential}(\alpha^*)\}$; the choice between h-likelihoods based on $v \equiv \log(u)$ and on $v \equiv u^*$ cannot be made on grounds of additivity.

(ii) The use of h-likelihood or of extended quasi-likelihood leads typically to inconsistent estimators, even after the various bias adjustments. Consistency, of course, is no guarantee of quality. But, if an estimator has persistent bias of order $O(1)$ as the amount of relevant data increases, there are some situations where the results will be seriously misleading, on account of short confidence intervals which systematically fail to contain the true value. This may be of no consequence in a *particular* application, but it does indicate a need for moderation in claims that are made for universality of the methods. Various unqualified assertions made in this paper, such, as 'statistically valid inferences', 'statistically efficient', 'always gives statistically sensible estimators' and 'interval estimators maintain the nominal levels', should not be taken to apply universally.

(iii) The treatment of binary responses as a special case is unattractive relative to the apparent unity of the rest of the scheme, and it is perhaps misguided: it appears to be a reaction to the criticisms that were made of Lee and Nelder (1996), and as such it misses the point that the difficulties such as persistent bias are not specific to binary responses but can occur whenever the information per random effect is small.

In summary: the h-likelihood idea which lies at the core of Lee and Nelder's elegant algorithmic approach is not well defined, and estimation using h-likelihoods is not universally well behaved. The arbitrary dependence of h-likelihood on how a model is written is very unlikelihood like, and the properties of an estimation scheme that is based on h-likelihood need to be checked case by case when the information per random effect is not large. The notion of a single, computationally feasible, inferential framework for a very large class of models has much appeal, but the authors' position at the very end of this paper (Section 7) is more

realistic: h-likelihood may not always give good results, needs its properties to be carefully assessed in each application and may need to be modified when problems are found.

With much regret that I have no time left to discuss the paper's more adventurous modelling aspects, I am pleased to second the vote of thanks.

The vote of thanks was passed by acclamation.

Robert A. Rigby (*London Metropolitan University*)

Random effects have previously been included in the scale parameter of a distribution by Rigby and Stasinopoulos (2005), so I would like to compare the authors' 'double hierarchical generalized linear model' (DHGLM) with the 'generalized additive model for location, scale and shape' (GAMLSS) of Rigby and Stasinopoulos (2005).

The GAMLSS is specified by

$$Y|\gamma's \sim D(\mu, \sigma, \nu, \tau)$$

where D is any distribution and

$$\begin{aligned} g_1(\mu) &= \eta_1 = X_1\beta_1 + Z_1\gamma_1, \\ g_2(\sigma) &= \eta_2 = X_2\beta_2 + Z_2\gamma_2, \\ g_3(\nu) &= \eta_3 = X_3\beta_3 + Z_3\gamma_3, \\ g_4(\tau) &= \eta_4 = X_4\beta_4 + Z_4\gamma_4. \end{aligned}$$

Here μ, σ, ν and τ are distribution parameters typically representing location, scale, skewness and kurtosis respectively, the γs are random effects and $\gamma_k \sim N(0, G_k)$ with covariance matrix G_k which may depend on random-effects hyperparameters λ_k, for $k = 1, 2, 3, 4$.

The DHGLM in a similar notation is specified by

$$Y|\nu, b \sim \mathrm{GLM}\{\mu, \phi\, V(\mu)\}$$

where

$$\begin{aligned} g_1(\mu) &= \eta_1 = X_1\beta_1 + Z_1 v, \\ g_2(\phi) &= \eta_2 = X_2\beta_2 + Z_2 b. \end{aligned}$$

Here v and b are random effects.

The first major difference between the models is that the GAMLSS allows *any* distribution for Y given the random effects (including positively or negatively skew and/or leptokurtic or platykurtic continuous or discrete distributions). The DHGLM uses a GLM distribution, which, in my opinion, is very restrictive. The GLM family does not include negatively skew, platykurtic or truncated distributions, or, for example, any of the continuous distributions beta, Gumbel, Weibull, logistic, exponential power, generalized inverse Gaussian, skew normal or skew t, or any of the discrete distributions negative binomial type II, Sichel, Delaport or zero-inflated negative binomial. All these are available in the GAMLSS.

Even within the GLM family, for most variance functions $V(\mu)$ the GLM distribution either does not exist or has a probability (density) function that is mathematically intractable. The authors approximate such distributions by using extended quasi-likelihood. However, extended quasi-likelihood is not a proper distribution as it does not integrate (or sum) to 1. Furthermore the integrand generally cannot be obtained explicitly (requiring numerical integration), varies between observations and crucially depends on the parameters of the model and so should be included in the likelihood to be maximized, requiring a numerical optimization procedure rather than the authors' iterative reweighted least squares procedure.

A second difference between the models is that the GAMLSS allows modelling of all parameters of the distribution of Y given the random effects in terms of explanatory variables and random effects. The current implementation of the GAMLSS allows up to four parameters for this distribution (μ, σ, ν and τ), allowing changes in shape to be modelled explicitly. In contrast, the DHGLM models the location and scale, μ and ϕ, only.

In conclusion, I believe that the DHGLM may be suitable for data that fall within its model range, but that the GAMLSS is a more flexible and transparent model.

I congratulate the authors on a stimulating paper.

D. M. Stasinopoulos (*London Metropolitan University*)

It is nice to see that the authors always have something new to add to their original hierarchical generalized linear model formulation. They first introduced a random effect in the mean; they now add a random effect in the variance (which enables them to fit certain kurtotic distributions). Also smoothing is now included. Of course, other frameworks such as the generalized additive model for location, scale and shape (GAMLSS) (Rigby and Stasinopoulos, 2005) do both these.

My main concern lies with a part of their original formulation of the hierarchical generalized linear model. More precisely I am concerned about the use of extended quasi-likelihood (EQL), for the conditional distribution of the response y, because it is not a proper distribution.

For example, in the epileptic data example of Section 3.2, the authors propose a model with the conditional distribution of y defined as an exponential family distribution with the property that $V(y) = \phi\mu$. This distribution does not exist, so the authors use EQL. Let us compare their EQL distribution with the negative binomial type II distribution which has the same variance–mean relationship, i.e. variance of y proportional to μ.

Fig. 5 shows the two distributions both with $\mu = 10$ and $\phi = 3$. The difference in shape is evident. The EQL distribution has been standardized to sum to 1 but in reality it is not a proper distribution since it sums to 1.0383. This roughly means that reported maximized likelihoods could be inaccurate by almost 4%. (The mean of the EQL is 9.874, approximately μ, and the variance is 30.105, approximately $\phi\mu$). The same problem arises with overdispersed binomial data as shown in Fig. 6 with $N = 20$, $p = 0.5$ and $\phi = 3$ where the EQL probabilities sum to 1.057 and the distributions, EQL and beta–binomial, are very different. My feeling is that EQL should now take its rightful place in history. In this age there are other frameworks which allow us to fit proper distributions. Jim Lindsey has done much work on fitting overdispersed distributions (which sum to 1) and his software is available from `http://popgen0146uns50.unimaas.nl/~jlindsey/`.

The GAMLSS software (in the R language) allows rapid fitting of different distributions and can be downloaded from CRAN. The GAMLSS has a 200-page manual available from `http://www.londonmet.ac.uk/gamlss/`.

I congratulate the authors. I always find their papers interesting and challenging.

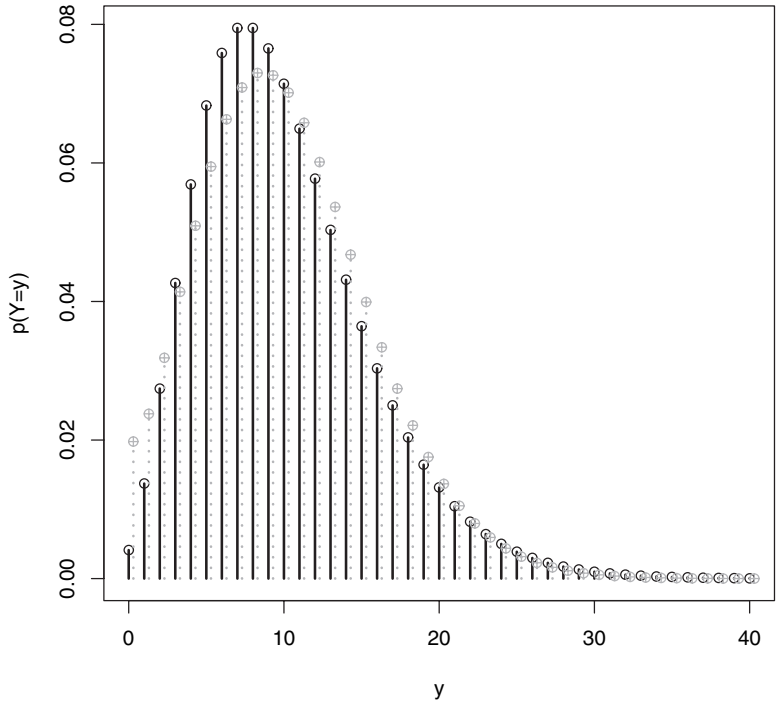


Fig. 5. Negative binomial type II (|) and EQL overdispersed Poisson distributions (⋮) with $\mu = 10$ and $\phi = 3$

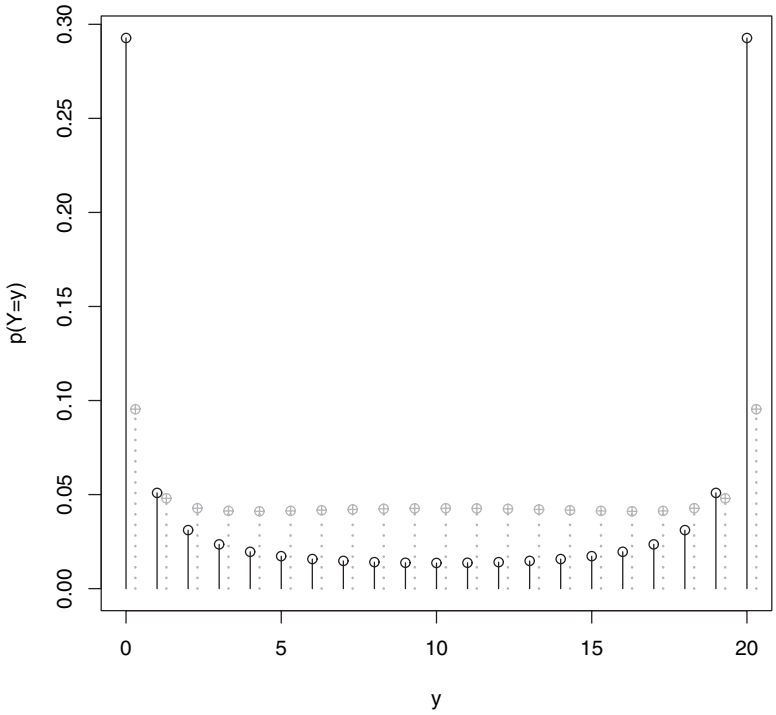


Fig. 6. Beta–binomial (|) and EQL overdispersed binomial distributions (⋮) with $N = 20$, $p = 0.5$ and $\phi = 3$

Roger Payne (*Rothamsted Research, Harpenden*)

I believe that this is a very important paper for statistical analysis, which shares many of the characteristics of the ground breaking paper by Nelder and Wedderburn (1972) that first defined generalized linear models (GLMs). From the perspective of an applied statistician who might wish to make use of the techniques, the following seem to me to be the key features.

(a) Double hierarchical GLMs follow a unifying approach, bringing a wide range of models together within a single framework. This empowers people, giving them the flexibility to select a good model, rather than the one that is least inappropriate. This is exactly what the earlier GLM framework achieved, resulting in a noticeable improvement in the quality of statistical analysis.

(b) They support (and encourage the recognition of) multiple sources of error variation. This is a long-standing interest at Rothamsted, of course, which dates back to Fisher and Yates's development of the split-plot design, and has continued for example with the definition and further refinement of the concept of general balance; see Nelder (1965) and also for example Payne and Tobias (1992) and Payne (2004).

(c) They were motivated by John Nelder's experiences with agricultural and biological researchers (sadly, an area where statistics now seems to be in decline) but, like GLMs, they are relevant wherever statistics are used.

(d) They are fitted by a computationally efficient algorithm. So we are encouraged to analyse our data interactively, and to find the right model rather than simply stopping with relief when we find one that converges. The timings that are given in the paper are impressive, but I can report that I am working with Youngjo and John on a reimplementation of the algorithms, which is giving about a tenfold improvement in speed, as well as enabling us to form tables of predicted means (see Lane and Nelder (1982)).

(e) The fact that each hierarchical GLM is made up from two interlinked GLMs means that we have access to a familiar *repertoire* of model checking techniques and can thus base our choice of error distributions on the data rather than on prejudice or limitations of software. This is particularly important given the richness of the class of models that we can now fit, and it will be an interesting task to build up the experience to determine when each of these models will be most effective!

Stephen Senn (*University of Glasgow*)

The ability to model mean and dispersion together is important. For example, all current standard approaches to meta-analysis, including random-effects approaches, treat the observed within-trial precision as if it were a known quantity (Senn, 2000). This, as was already known to Yates and Cochrane (Yates and Cochran, 1938), biases tests of significance and confidence intervals and leads to inefficient estimates. The double hierarchical generalized linear models (DHGLMs) of Lee and Nelder should prove extremely valuable in this context.

The authors should not underestimate, however, the importance of making these models easy to use. Most statisticians (alas) do not use Genstat® and even those, like me, who do, balk at using the *k*-system. Thus the efforts that Roger Payne and others have made to make HGLMs capable of easy implementation in Genstat®, whether in command or menu mode, have been crucial in making it possible for me to use Lee and Nelder's earlier HGLM (Lee and Nelder, 1996) work. I suspect that I will not be the only applied statistician who will be reliant on further progress at Genstat® to make implementation of DHGLMs a practical personal possibility. Lee and Nelder have built a fine mouse-trap but that does not mean that the world will beat a path to their door.

If I have one gripe with this paper, it is the use of the tired and unconvincing epilepsy example of Thall and Vail (1990). A crossover trial has had the second period data thrown away to make it a suitable guinea-pig for analysis. The patient numbers run (with 17 missing numbers) from 101 to 147 and then again (with nine missing numbers) from 201 to 238. This immediately suggests a two-centre trial with many drop-outs. Checking this point is difficult, since the reference in Thall and Vail (1990) is incorrect. A paper with a different title, but the same first author (Leppik *et al.*, 1987), also describing a crossover comparing progabide and placebo in 59 epilepics, appeared in *Neurology* in 1987 (not 1985) and is almost certainly the study on which the epilepsy data are based. Thall and Vail (1990) had been cited 82 times by October 2005: presumably nobody has bothered to find the source. If they had, they would have discovered a trial in two centres with two strata (by type of seizure) per centre (Leppik *et al.*, 1987).

In my view the analysis of data as four visits is pointless and we might as well analyse the totals. Leppik *et al.* (1987), using all the data of the original crossover trial, found no convincing evidence of a treatment effect and I am suspicious of any analyses of the first-period data only, including those of Lee and Nelder and Thall and Vail (1990), that do. Fitting total seizures as a function of centre, age and base-line seizure in addition to treatment using either Poisson regression and allowing for overdispersion or a negative binomial model, or using the square root of the number of seizures in a linear model, I find no convincing evidence of a treatment effect. In my view nothing of any use can be decided about the relative value of any methods in analysing this data set. It is time that it was retired.

I congratulate the authors on their latest advance in developing tools for others and wish them every success in making them easily usable.

The following contributions were received in writing after the meeting.

William J. Browne (*University of Nottingham*) **and Harvey Goldstein** (*University of Bristol*)

We congratulate Lee and Nelder on an impressive synthesis of many statistical models into an extended family from a frequentist perspective. We have several comments about various aspects of the paper. Firstly we have previously developed models for structuring the dispersion in Gaussian response models both from a likelihood (Goldstein, 1986) and a Bayesian perspective (Browne *et al.*, 2002) and implemented such approaches in the MLwiN software package (Rasbash *et al.*, 2004). We allow the dispersion to vary with the value of predictor variables in a similar way to Lee and Nelder's hierarchical generalized linear models with structured dispersion in examples 3.1 and 3.2. In these examples the motivation of using random effects on the dispersion appears to be to account for outlying units. An alternative approach of course would be to account for just these outlying units by using dummy variable fixed effects in the dispersion generalized linear model, rather than resorting to random effects, and simply to fit a hierarchical generalized linear model with structured dispersion. Have the authors considered this approach and how does it compare on their examples?

Secondly, we wish to comment on the *h*-likelihood approach and the adaptation that Lee and Nelder have used for binary data. We are impressed that the biases that were originally found for binary data appear to have been reduced by their adjusted profiling functions. In previous work (Browne, 1998) we have considered the point estimate biases of marginal quasi-likelihood, penalized quasi-likelihood and Markov chain Monte Carlo (MCMC) methods with non-informative priors for 500 simulated three-level data sets based on a real data set from Rodríguez and Goldman (1995). The underlying model for the simulations is

Table 10. Point estimates (with Monte Carlo standard errors) for parameters in a three-level logistic regression model based on 500 simulations

Parameter (true)	*Estimates from the following methods:*			
	1st-order marginal quasi-likelihood	*2nd-order penalized quasi-likelihood*	*MCMC, gamma prior*	*MCMC, uniform prior*
β_0 (0.65)	0.474 (0.01)	0.612 (0.01)	0.638 (0.01)	0.655 (0.01)
β_1 (1.00)	0.741 (0.01)	0.945 (0.01)	0.991 (0.01)	1.015 (0.01)
β_2 (1.00)	0.753 (0.01)	0.958 (0.01)	1.006 (0.01)	1.031 (0.01)
β_3 (1.00)	0.727 (0.01)	0.942 (0.01)	0.982 (0.01)	1.007 (0.01)
σ_v^2 (1.00)	0.550 (0.01)	0.888 (0.01)	1.023 (0.01)	1.108 (0.01)
σ_u^2 (1.00)	0.026 (0.01)	0.568 (0.0l)	0.964 (0.02)	1.130 (0.02)

$$y_{ijk} \sim \text{Bernoulli}(p_{ijk})$$

with

$$\text{logit}(p_{ijk}) = \beta_0 + \beta_1 x_{1ijk} + \beta_2 x_{2jk} + \beta_3 x_{3k} + u_{jk} + \nu_k$$

where $u_{jk} \sim N(0, \sigma_u^2)$ and $\nu_k \sim N(0, \sigma_\nu^2)$.

We found that there was a bias with first-order marginal quasi-likelihood that was substantially reduced when second-order penalized quasi-likelihood was used (Goldstein and Rasbash, 1996) and further reduced using MCMC sampling with alternative choices of priors. The biases were greatest for the variance parameters (Table 10) and we wonder how Lee and Nelder's methods perform for these parameters.

Finally we are appreciative of the effort that Lee and Nelder have made to extend their framework to yet further models in this paper. Of course the models that they suggest can also be fitted by using MCMC algorithms in a Bayesian framework and generally this approach is easier to implement and extend. Do the authors believe that like MCMC sampling their approach can be adapted still further to incorporate many sets of correlated random effects, missing data and measurement errors?

Joan del Castillo (*Universitat Autònoma de Barcelona*)

The authors are to be congratulated for developing a further extension of hierarchical generalized linear models (HGLMs), allowing joint modelling of the mean and dispersion with random effects in both structures. GLMs and random-effects models are very common in statistical practice. However, only normal random effects are usually considered in the standard software, even bearing in mind that the normality assumption is vulnerable to outliers. Certain misunderstandings of HGLM models may lead to the impression that only conjugate families are possible within such a model. It is now made clear that this assumption is not required at all, and that non-normal random effects should be considered in more general situations.

The double hierarchical generalized linear models (DHGLMs) are especially welcome in financial economics and mathematical finance. Stochastic volatility is the main concept that is used in these fields to deal with time-varying volatility in financial markets. Volatility, the main objective, is a random effect in the dispersion structure. Hence, the double hierarchical model is the tool facilitating the inclusion of stochastic volatility models in the GLM context, as Section 2.2 shows. The h-likelihood method for DHGLMs, which avoids the integration that is necessary to obtain marginal likelihood, can be used to estimate fixed and random parameters for a large class of parametric Lévy processes, as considered in Eberlein and Keller (1995) or Barndorff-Nielsen and Shephard (2001). In particular, for variance gamma Lévy processes, which were introduced by Madan and Seneta (1990), no analytic formula for the marginal probability density function was known at the time of the paper's publication. Consequently, onerous numerical integration methods were used.

The DHGLMs enable many statistical problems to be modelled in a unified manner. The h-likelihood estimation method provides an extension of the Fisher likelihood framework, which allows inference from models with both fixed and random parameters. It is therefore important to consider this alternative

method, even if it might be subject to small amounts of bias. I have some questions for the authors concerning the requisites of careful guidelines for use. When might bias be severe? Do second-order corrections need to be used for financial models?

Mohand Feddag (*University of Warwick, Coventry*)
To familiarize myself with the approach that is taken in this paper, I conducted a small simulation experiment. The results do not seem to agree with some remarks that are made in the paper. I compared the method of maximum adjusted profile h-likelihood (MAPHL) with maximum marginal likelihood (MML), for a binomial–beta hierarchical generalized linear model with the random-effect distribution assumed known. Let $\{y_{ij} : i = 1, \ldots, n; j = 1, 2\}$ be such that

$$y_{ij}|u_i \sim \text{Bernoulli}(p_{ij}),$$

$$\text{logit}(p_{ij}) = \beta_j + \text{logit}(u_i),$$

$$u_i \sim \beta(\alpha_1, \alpha_2).$$

We assume that $\beta_1 = 0$ and that α_1 and α_2 are given. Interest is in the estimation of the parameter β_2 by the MAPHL and MML methods. The h-likelihood is

$$\begin{aligned} h &= l(\beta; y|v) + l(\alpha_1, \alpha_2; v) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{2}[y_{ij}(\beta_j + v_i) - \ln\{1 + \exp(\beta_j + v_i)\}] + \alpha_1 \sum_{i=1}^{n} v_i - (\alpha_1 + \alpha_2)\sum_{i=1}^{n} \ln\{1 + \exp(v_i)\} - n \ln\{B(\alpha_1, \alpha_2)\}. \end{aligned}$$

The first and second partial derivatives with respect to v_i are

$$\frac{\partial h}{\partial v_i} = \sum_{j=1}^{2}\left\{ y_{ij} - \frac{\exp(\beta_j + v_i)}{1 + \exp(\beta_j + v_i)} \right\} + \alpha_1 - (\alpha_1 + \alpha_2)\frac{\exp(v_i)}{1 + \exp(v_i)},$$

$$\frac{\partial^2 h}{\partial v_i^2} = -\sum_{j=1}^{2} \frac{\exp(\beta_j + v_i)}{\{1 + \exp(\beta_j + v_i)\}^2} - (\alpha_1 + \alpha_2)\frac{\exp(v_i)}{\{1 + \exp(v_i)\}^2}.$$

The MAPHL approach estimates β_2 by maximizing

$$p_v(h) = \left[h - \frac{1}{2} \ln\left[\det\left\{ \frac{D(h, v)}{2\pi} \right\} \right] \right]_{\hat{v}},$$

where $D(h, v) = -\partial^2 h / \partial v^2$ and $\hat{v}$ solves $\partial h / \partial v = 0$. This is achieved via a two-step iteration: for given β_2 estimate v by maximizing h, and for given v maximize $p_v(h)$ for β_2.

Table 11 shows the empirical mean and standard deviation of computed MAPHL and MML estimates for 500 samples with $\alpha_1 = 1$, $\alpha_2 = 2$ and various values of β_2 and n. The MAPHL estimator appears to exhibit substantial bias, whereas MML behaves much better. This seems to be at odds with some assertions that are made in the paper (e.g. Section 4) concerning the performance of h-likelihood with binary data. It is entirely possible, of course, that I have made an error in interpreting details of the h-likelihood method,

Table 11. Binomial–beta model: means (and standard deviations in parentheses) of the MML and MAPHL estimators in 500 simulated samples

Method	*Results for the following values of* β_2 *and* n*:*					
	$\beta_2 = 1$			$\beta_2 = 2$		
	$n = 50$	$n = 100$	$n = 200$	$n = 50$	$n = 100$	$n = 200$
MML	1.02 (0.39)	1.03 (0.27)	1.00 (0.19)	2.04 (0.43)	2.02 (0.29)	2.00 (0.20)
MAPHL	0.76 (0.37)	0.77 (0.25)	0.65 (0.17)	1.81 (0.41)	1.76 (0.26)	1.63 (0.18)

or in programming it; unfortunately I could not easily check the calculations by using the authors' own programs because they require proprietary software to which I do not have access. In their response I hope that the authors might be able to shed light on the apparent discrepancy.

Il Do Ha (*Daegu Haany University, Gyeongsan*)
I welcome this paper, which introduces a further hierarchical generalized linear model (HGLM) framework to allow random effects in the linear predictors of both the mean and dispersion. In particular, the double hierarchical generalized linear models (DHGLMs) can unify various models and lead to robust inference against outliers or misspecification of the fatness of tails. In my view the introduction of DHGLMs in frailty models provides a very useful and powerful framework in survival analysis. The heterogeneity of hazards between clusters can be modelled by introducing frailties in the hazard (Aalen, 1988; Hougaard, 2000). Similarly, the heterogeneity of dispersions between clusters, which can in turn describe abrupt changes between recurrent or multiple-event times on the same cluster, can be modelled by introducing frailties in the dispersion (e.g. the variance) of frailty. As HGLMs (with structured dispersions) lead to (dispersion) frailty models (Noh *et al.*, 2006) DHGLMs allow a useful extension of frailty models. Furthermore, in frailty models many researchers including Xue (2001) and Ha and Lee (2005) have pointed out that the estimation of dispersion parameters can be vulnerable against the misspecification of the frailty distribution. As in Section 2.1, the use of heavy-tailed distributions for frailties would give robust inferences for both fixed and dispersion parameters: for ascertainment adjustment see Noh *et al.*, (2005). I believe that the use of the h-likelihood will provide statistically and numerically efficient estimation in many statistical areas. This is certainly true in frailty models (Ha and Lee, 2005).

Donghoh Kim (*Hongik University, Seoul*) **and Hee-Seok Oh** (*Seoul National University*)
It is a pleasure to comment on this interesting paper. The major contributions of the paper are to introduce a class of an extension of generalized linear models with random effects for both the mean and dispersion, and to use h-likelihood for unifying the framework of these extended models while providing an efficient algorithm for fitting.

Before discussing double hierarchical generalized linear models (DHGLMs), we would like to mention our successful experience with HGLMs as an imputation principle in wavelet regression problems. When missing values are present, most existing wavelet regression methods cannot be directly applied to recover the true function. Suppose that the complete data $\mathbf{y}_{\text{com}} = (\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}})$ are from normal distribution $N(\mathbf{f}, \sigma^2 I_n)$, where $\mathbf{y}_{\text{obs}} = (y_1, \ldots, y_k)$ and $\mathbf{y}_{\text{mis}} = (y_{k+1}, \ldots, y_n)$ denote observed and missing values respectively. Let $\mathbf{f} = (f_1, f_2, \ldots, f_n)$, where $f_i = (\mathbf{W}^{\mathrm{T}}\boldsymbol{\theta})_i$, $\mathbf{W}$ is the orthogonal wavelet operator and $\boldsymbol{\theta}$ denotes wavelet coefficients. The goal is to estimate the function $\mathbf{f}$ when there are missing values. For this, we employ the h-likelihood principle of treating the missing values as random effects, and then we consider the *penalized log-likelihood* of the complete data:

$$P_{\text{com}} = H_{\text{obs}} + H_{\text{mis}} - \lambda\, q(\boldsymbol{\theta}),$$

where

$$H_{\text{obs}} = -\frac{1}{2\sigma^2}\sum_{i=1}^{k}(y_i - f_i)^2,$$

$$H_{\text{mis}} = -\frac{1}{2\sigma^2}\sum_{i=k+1}^{n}(y_i - f_i)^2,$$

λ is the thresholding values for wavelet shrinkage and $q(\boldsymbol{\theta})$ is a penalty function. The maximization of this penalized likelihood is equivalent to the minimization of

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n}(y_i - f_i)^2 + \lambda\, q(\boldsymbol{\theta}).$$

Then the missing data $\mathbf{y}_{\text{mis}}$ can be imputed by solving the score equations

$$\frac{\partial P_{\text{com}}}{\partial y_i} = -\frac{1}{\sigma^2}(y_i - f_i)$$

$(i = k+1, \ldots, n)$. It results in $\hat{y}_{\text{mis},i} = f_i$. Note that our imputation $\hat{y}_{\text{mis},i}$ of missing data is the best unbiased predictor

$$E(\hat{y}_{\text{mis},i}|\mathbf{y}_{\text{obs}}) = E(\hat{y}_{\text{mis},i}) = f_i.$$

Since this derivation does not use any Monte Carlo simulation or approximations, the proposed approach provides a fast and efficient algorithm for wavelet regression with missing data.

We believe that DHGLMs can be applied to the imputation principle for heavy-tailed distributions and correlated cases. DHGLMs broaden the linear model with random effects not only for the mean but also for dispersion, which allows flexible modelling. Thus by treating the missing values as the random-mean part and modelling the random-dispersion part properly, DHGLMs will provide a useful imputation tool to cope with correlation and heavy tailedness.

Andrew B. Lawson and Walter W. Piegorsch (*University of South Carolina, Columbia*)

We congratulate Professor Lee and Professor Nelder on a stimulating paper; their basic premise that generalized linear models can be extended by incorporating hierarchical features in both the mean and the dispersion components is underutilized in the literature, despite its potential to expand greatly the class of models that are available to the data analyst. Their specific incorporation of random effects into the mean *and* dispersion components widens the model framework in a novel and intriguing fashion, allowing for even broader classes of models and, as Lee and Nelder's various examples illustrate, opening the possibility of more enhanced model formulations, more robust model fits and more pertinent interpretations of the model parameters. We warmly welcome such efforts in statistical modelling.

We were confused, however, by a few minor issues. First, we were uncertain how random effects in the λ-component can have parameter estimates that are insensitive to misspecification of the distribution of random effects (Section 2). Our own experience—albeit not with models as complex as double hierarchical generalized linear models (DHGLMs)—suggests that model misspecification can affect a random-effect term in a powerful and often undesirable fashion. We encourage the authors to expand on this assertion.

Second, we caution readers not to interpret the statement

> '... DHGLM has the smallest standard error estimates, reflecting the gain in information from having proper dispersion modelling'

(at the end of Section 3.2) as meaning that a method yielding smaller standard errors is necessarily better in any given situation (as we at first did). A given model whose fit yields smaller standard errors than another competitor may give more powerful inferences, but of course without knowledge of the truth the data analyst cannot guarantee that smaller standard errors actually indicate a more correct model fit.

Finally, as many before us have noted, 'hierarchical' does not always mean 'Bayesian' and we caution readers against making such an automatic connection. Indeed, Lee and Nelder do note that the likelihood method does not use priors. But, if we were to adopt a Bayesian perspective here, the method could be viewed as employing uniform priors and, as a result, yielding only a small fraction of the information that is provided by posterior sampling. By opening their model formulation to a wider class of prior distributions, the authors could broaden the extent and applicability of their DHGLMs.

In any case, we once again commend Professor Lee and Professor Nelder for their intriguing paper. We encourage them and our many modeller colleagues to incorporate predictor variables and random effects across the various levels of a model's hierarchy wherever the opportunity presents itself.

Geert Molenberghs (*Hasselt University, Diepenbeek*) **and Geert Verbeke** (*Katholieke Universiteit Leuven*)

Hierarchical data are very common. Over the last three decades, various approaches have been suggested (Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993) and subsequently refined. Lee and Nelder (1996, 2001a, b, 2005a, b) have strongly contributed to this area. In spite of communality there are important differences also, in model formulation and inferential route taken. Within the marginal likelihood framework, we distinguish between three broad families:

(a) approximation to the integrand (e.g. Laplace transforms),
(b) approximation to the data (penalized quasi-likelihood and marginal quasi-likelihood) and
(c) approximation of the integral (numerical integration) (Molenberghs and Verbeke, 2005).

As the authors state, all of these suffer from drawbacks. Numerical integration can be highly accurate, even for binary data, but the computational requirements often are prohibitive. Penalized quasi-likelihood and marginal quasi-likelihood methods can be severely biased, especially for binary outcomes. Not surprisingly, this is also so where the initial h-likelihood formulation performed poorest. In the light of this,

the criticism that penalized quasi-likelihood, marginal quasi-likelihood and h-likelihood have received was perhaps less evenly distributed than it ought to have been.

Therefore, the recent work of the authors is very welcome. It refines h-likelihood and gives it its proper place as a *third way* in between marginal likelihood and Bayesian methodology. Admittedly, the method is less straightforward to apply than, for example, ordinary least squares regression, but this is true for all its competitors also and derives from the complexity of data and likelihood. A potential user should carefully read the method's manual, e.g. that it is important to distinguish between mean and variance component estimation and that such correction functions as $p_{v,\beta}(h)$ should be used whenever appropriate. Such seemingly *ad hoc* aspects abound throughout complex data analysis and should not be considered a disadvantage of the method proposed.

Although modelling strategies for the variance structure are not new, ranging from conventional and well-understood structures stemming from random effects in linear mixed models, factor analytic structures and antedependence structures, to proposals that have been made by Pourahmadi (2000) and others, the current proposal enjoys synthetic powers. Not only does it unify several modelling families that are often considered distinct, it also places overdispersion, variance and covariance modelling, including flexibility of skewness and kurtosis, under one umbrella. Clearly, some of these modelling ideas are logically independent of the h-likelihood framework, which is the authors' inferential choice of preference. This gives wide applicability to the paper's modelling ideas.

Kelvin K. W. Yau (*City University of Hong Kong*)

I congratulate the authors for their contribution on advocating a class of double hierarchical generalized linear models (DHGLMs). I can only add a few remarks to this interesting and important study.

(a) In the full DHGLM specifications (Section 2), the number of fixed effect parameters ($p_1 + p_2 + p_3 + p_4$) has been increased considerably when compared with a conventional generalized linear mixed model (GLMM) (in which the number of parameters is $p_1 + 1$). I wonder whether there are any difficulties in estimation owing to the increased number of parameters to be estimated. Furthermore, in most practical situations, the same model matrix will be used in the four components, i.e. $X = G_m = G = G_d$. Will there be potential identifiability problems because of the use of the same design matrix in the four components?

(b) In Section 2.1, by introducing a random component in the dispersion model, a heavy-tailed distribution (here the multivariate t-distribution) will be induced for the random effects in the mean model, and this is proposed to achieve robust estimation against outliers. Alternatively, Yau and Kuk (2002) proposed the use of the robust log-likelihood in the GLMM to limit the influence of large random-effect terms and error terms in the adjusted dependent variable for the mean model. Both proposals are plausible approaches to achieve robust estimation against outliers.

(c) In the application of Poisson HGLMs and the conjugate Poisson DHGLM to the data on epileptics (Section 3.2), it is found from Table 3 that the effect of visit (β_V) is not consistent throughout. Previous studies (Lee and Nelder, 1996; Yau and Kuk, 2002) have identified the third observation of patient 227 as the only outlier. By limiting the influence of outlying observations, the data are reanalysed by using the robust GLMM approach (Yau and Kuk, 2002). In particular, two Poisson GLMMs are considered:

$$\eta_{ij} = x'_{ij}\beta + a_i,$$

$$\eta_{ij} = x'_{ij}\beta + a_i + b_{ij},$$

where x_{ij} is a vector of explanatory variables that are associated with the ijth observation and a_i and b_{ij} are normally distributed and represent respectively the subject level and unit level random effects. The extra random term in the second model can be viewed as modelling overdispersion at the unit level, which is in principle similar to the situation in the quasi-HGLM and the HGLM with structured dispersion for modelling extra variation. The model fitting results are given in Table 12. It is interesting that the effect of the visit variable becomes not significant when shifting from a one-level to a two-level random-effect model.

Keming Yu and Rogemar Mamon (*Brunel University, Uxbridge*)

Double hierarchical generalized linear models (DHGLMs) are indeed a big family of models and h-likelihood is a unified approach for fitting DHGLMs. We just point out that a class of double semiparametric hierarchical linear models (DSHLMs), in which random effects can also be specified for both the mean

Table 12. Robust GLMM fits to the epilepsy data†

Parameter	*Estimates for the one-level random-effect model* $\eta_{ij}=x'_{ij}\beta+a_i$ *and the following methods:*		*Estimates for the two-level random-effect model* $\eta_{ij}=x'_{ij}\beta+a_i+b_{ij}$ *and the following methods:*	
	Robust REML I	*Robust REML II*	*Robust REML I*	*Robust REML II*
β_0	−1.25 (1.28)	−1.24 (1.31)	−1.27 (1.28)	−1.27 (1.31)
β_B	0.84 (0.14)‡	0.84 (0.15)‡	0.85 (0.14)‡	0.85 (0.15)‡
β_T	−0.94 (0.43)‡	−0.94 (0.44)‡	−0.93 (0.43)‡	−0.93 (0.44)‡
β_{T*B}	0.34 (0.22)	0.34 (0.23)	0.34 (0.22)	0.34 (0.22)
β_A	0.47 (0.38)	0.47 (0.38)	0.47 (0.38)	0.46 (0.38)
β_V	−0.32 (0.11)‡	−0.32 (0.11)‡	−0.25 (0.17)	−0.25 (0.17)
var(a_i)	0.276	0.294	0.242	0.255
var(b_{ij})			0.129	0.132

†Standard errors are given in parentheses.
‡Significant at the 5%-level.

and the dispersion, can be fitted by a least squares (LS) approach without resorting to likelihood, which is usually unknown or poorly specificed. The LS approach is also simple and fast in computing. In fact, the DSHLMs could be defined as

$$Y_{ij}=X'_{ij}\beta+f(t_{ij})+\exp\{X'_{ij}\gamma+g(t_{ij})\}\varepsilon_{ij},$$

where β and γ are the level 1 parameters which may depend on other variables being predicted from level 2 variables and so on. The functionals f and g are unknown but could be real functions or random components, and ε_{ij} are random errors with $E(\varepsilon_{ij})=0$ and $V(\varepsilon_{ij})=\sigma^2$. Instead of the likelihood approach or the h-likelihood technique, we could implement a double LS regression technique to fit the DSHLMs. For example, we can estimate β and f by using P-splines and the LS approach based on observations on (Y, X, t); then we can estimate γ and g by using P-splines and the LS algorithm again but based on observations on the logarithm of the residual squares from the previous mean fitting.

Both the h-likelihood technique and the proposed alternative method of model fitting that is described above can then be applied to, say, a financial data set. On the basis of forecasting performance such as the quality of k-step-ahead predictions, $k>0$, or possibly by means of other bench-marks, we can compare the results that are generated by the h-likelihood and double LS regression. The effect and performance of DGLMs via h-likelihood implementation may be explored further, at least within the finance or econometrics context, by performing a comparison of results with those that will be generated by other competing estimation techniques of stochastic volatility models that have been suggested by others. These include, among others, the simulated method of moments of Duffie and Singleton (1993), the indirect inference of Gourieroux *et al.* (1993), the efficient methods of moments of Gallant and Tauchen (1996) and the simulated likelihood methods of Sandmann and Koopman (1998).

Zhong-Zhan Zhang (*Beijing University of Technology*)

I congratulate the authors on the new development of hierarchical generalized linear models. Hierarchical generalized linear models have been studied by many researchers, as can be seen in the references that are listed in the paper. The class of double hierarchical generalized linear models is a meaningful extension of hierarchical generalized linear models and includes a great variety of models. The authors propose a motivating framework for modelling mixed dispersions and give an easily implemented and numerically efficient algorithm.

There are few methods which could predict the cluster-specific random effects for generalized linear models well. h-likelihood inference was proposed heuristically. For Poisson–gamma-type models, we have found that the estimators of fixed effects can have a convergence rate of the iterative logarithm law (Xia and Zhang, 2005a). For general multiplicative models, whenever the random effects can be predicted unbiasedly, the estimators of fixed effects that are obtained would be consistent (Chen *et al.*, 1999); otherwise, the estimating equation that is obtained from h-likelihood might have a bias. Theoretically, the

estimators through marginal likelihood could be better; however, when the integral is difficult to compute precisely, the estimator that is obtained through approximations could also be biased (Lin and Breslow, 1996), especially for small and moderate sample sizes. We have found by simulations that the algorithm is very numerically efficient, and the estimates that are obtained are quite good, even for the cases in which several random effects are combined (Xia and Zhang, 2005b).

The double hierarchical generalized linear model class includes a generalized linear model part with a random effect for dispersion, so it is more flexible. Modelling dispersion by using the generalized linear model has been employed in many fields; see, for example, Nelder and Lee (1991) in quality control and Smyth and Verbyla (1999) in environmental science. It has also been used in financial engineering with generalized autoregressive conditional heteroscedasticity models, as has been illustrated in the paper and elsewhere. From the theoretical point of view, the model class may be too rich to be covered by a unified theory. One difficulty is the properties of the predictions of random effects, especially for small sizes of clusters, which is more useful in medical practices. Ma (1999) studied unbiased prediction in a special model family. But model families can seldom lead to unbiased prediction, and the mean-square errors of the predictions are not available in general. For the properties of fixed effects, the key point is whether or not the estimating equations that are obtained are unbiased or asymptotically unbiased (Liang and Zeger, 1986). We look forward to more results of theoretical research.

The **authors** replied later, in writing, as follows.

We thank the discussants for raising many interesting points, and we regret that space restrictions do not allow us to reply to all of them.

The original Fisher likelihood framework is for a model class consisting of two types of object: observable random variables (data) and unknown fixed parameters. This paper concerns a general model class with an additional object, namely unobservable (or unobserved) random variables and an associated likelihood, namely the h-likelihood. We discuss

(a) the general model class,
(b) h-likelihood procedures as our inferential choice for such a model class and
(c) invariance, as raised by Firth.

Model aspects

We thank the discussants for noting that double hierarchical generalized linear models (DHGLMs) are useful for modelling meta-analysis, crossover trials and longitudinal studies (Senn), proportional hazards or non-proportional hazards frailty models with parametric and nonparametric base-line hazards (Ha and MacKenzie), covariance modelling (MacKenzie), quality control and environment science (Zhang), volatility in financial markets (Castillo, and Yu and Mamon), wavelet smoothing (Kim and Oh) and splines (Yu and Mamon). Molenberghs and Verbeke note that DHGLMs not only unify these families, but also include overdispersion, variance, covariance and the modelling of skewness and kurtosis under one umbrella. We agree with Professor Lawson and Professor Piegorsch that the extent and applicability of DHGLM formulation can be broadened by allowing a wider class of prior distributions.

Dr Rigby claims that the model class of their generalized additive models for location, scale and shape (Rigby and Stasinopoulos, 2005) is more general, which is not true. They consider some distributions with four or fewer parameters and claim to allow random effects in these parameters. If random effects are allowed in the kurtosis parameter they affect the eighth moment, so the original kurtosis parameters cannot be a kurtosis any more. Note that it is not true that 'the DHGLM uses a GLM distribution'. It is precisely by allowing random effects in the dispersion model that we avoid this restriction. Modelling of skewness and kurtosis is possible by allowing random effects in the mean and dispersion. By allowing various distributions we can systematically generate broad classes of new models. We are generally unhappy about modelling skewness and particularly about modelling kurtosis, because of the very large sample sizes that are required to achieve any usable accuracy. Tukey's 'rule of 5' is relevant here, namely that for every five readings for estimating the mean you need 5^2 for the dispersion, 5^3 for the skewness and 5^4 for the kurtosis. This is why we prefer to change the cumulant pattern by changing the variance function.

Professor Yau illustrates his robust analysis. Professor Lawson and Professor Piegorsch raise the question whether robust analysis is possible via statistical modelling. Noh and Lee (2005) showed that more general robust analysis is possible for the GLM class model via our modelling approach.

We must accept Senn's detailed criticism of our modelling of the epilepsy data. Our reason for using these data was to compare our fit with those of others.

Table 13. Three-level logistic regression model based on 500 simulations

Parameter (true)	*Results for model H(1)*			*Results for model H(2)*		
	Mean†	*Standard error*	*95%*	*Mean*†	*Standard error*	*95%*
β_0 (0.65)	0.634 (0.23)	0.22	94.4	0.657 (0.24)	0.24	94.8
β_1 (1.00)	0.986 (0.21)	0.21	93.8	0.994 (0.21)	0.21	95.2
β_2 (1.00)	0.983 (0.11)	0.10	93.2	1.007 (0.11)	0.11	94.8
β_3 (1.00)	0.980 (0.27)	0.26	94.2	1.013 (0.27)	0.28	94.6
σ_v^2 (1.00)	0.947 (0.19)			0.982 (0.22)		
σ_u^2 (1.00)	0.923 (0.14)			1.041 (0.16)		

†Standard deviations are given in parentheses.

h-likelihood approach

Professor Yu and Professor Mamon propose to use the log-normal fit for residual squares. The log-normal fit is very similar to the gamma fit with log-link, so it can be an alternative for obtaining splines for the dispersion by using a least squares approach.

Browne and Goldstein say that we use a 'frequentist perspective', but we do not; ours is a likelihood perspective. We do not accept the dichotomy that everyone is either a frequentist or a Bayesian. Our simulation results are in Table 13. Methods H(l) and H(2) are similar to the Markov chain Monte Carlo (MCMC) result with the gamma and uniform priors respectively. Their Monte Carlo standard error may correspond to our standard deviations divided by $\sqrt{500}$. The average of standard error estimates based on the Hessian matrix from h estimates the true standard deviation, so Wald confidence intervals maintain the required confidence level, especially with method H(2). Furthermore, the h-likelihood method has less bias in the estimation of dispersion parameters by using the restricted maximum likelihood methods $p_{v,\beta}(h)$. Restricted maximum likelihood adjustment is important when the number of fixed effects increases with sample size (Ha and Lee, 2005). We believe that the advantage of our method over MCMC methods lies in its efficiency; it is orders of magnitude faster than MCMC sampling. This means that model exploration becomes possible in a flexible way, which is surely always a good thing.

Breslow and Clayton (1993), Breslow and Lin (1995) and Lin and Breslow (1996) motivate the penalized quasi-likelihood (PQL) as a Laplace approximation and Lee and Nelder (1996, 2001a) also justified their adjusted profile h-likelihood $p_v(h)$ as a Laplace approximation. The difference between penalized likelihood and adjusted profile likelihood has not been clearly recognized. Rigby and Stasinopoulos (2005) used penalized likelihood as a criterion for fitting; this can give large biases in estimates. When we profile v by $\hat{v}$, $\hat{v}=\hat{v}(\beta)$ is a function of the parameters, so $\partial v(\beta)/\partial\beta$ cannot be ignored (Lee and Nelder, 2001a); however, the PQL-type estimator does ignore it. Feddag's maximum adjusted profile h-likelihood for binary matched pairs is even worse than our version of the PQL type. The h-likelihood estimator using $p_v(h)$ is very close to the marginal maximum likelihood (MML) estimator, obtained by numerical methods (Table 14). The advantage of using the h-likelihood approach is that it can be applied to various designs including crossed designs. Standard error estimates and confidence bounds are all satisfactory.

Firth recommends the use of composite likelihood when marginal likelihood is difficult to compute. For the Poisson variance function $V(\mu)=\mu$, let $y=\phi z$ for $\phi>0$, with

$$z \sim \text{Poisson}(\mu/\phi);$$

then $E(y)=\mu$ and $\text{var}(y)=\phi(\mu)$, but this distribution is not supported on the integer sample space $\{0, 1, \ldots\}$, but on $\{0, \phi, 2\phi, \ldots\}$. The use of this distribution for an inference is extended quasi-likelihood (EQL) for overdispersed Poisson distributions, which is equivalent to the use of a double-exponential family (Efron, 1986; Lee and Nelder, 2000). If this likelihood is used it gives a consistent estimator for the regression model for μ (Wedderburn, 1974), but biased estimation for ϕ unless $\phi=1$. However, Nelder and Lee (1992) showed that, although it can often give reasonable estimators for ϕ in finite samples, for some models it can give severely biased estimators (Ridout *et al.*, 1999); however, bias can be reduced substantially if properly treated (Lee, 2004). Both Rigby and Stasinopoulos complain about the use of EQL because of bias. For Poisson DHGLMs we can have heavy-tailed distributions by introducing random

Table 14. Binomial–beta model based on 500 simulations

Method	*Results for the following values of n:*								
	n = 50			*n* = 100			*n* = 200		
	Mean†	*Standard error*	*95%*	*Mean*†	*Standard error*	*95%*	*Mean*†	*Standard error*	*95%*
$\beta_2 = 1$									
MML	1.03 (0.40)	0.39	93.6	1.03 (0.27)	0.27	94.4	1.02 (0.19)	0.19	95.4
PQL type	0.75 (0.40)	0.38	91.0	0.81 (0.28)	0.25	89.2	0.80 (0.20)	0.19	88.4
$p_v(h)$	1.04 (0.40)	0.40	93.6	1.03 (0.28)	0.27	94.2	1.02 (0.19)	0.19	95.2
$\beta_2 = 2$									
MML	2.06 (0.41)	0.40	93.8	2.04 (0.28)	0.27	94.6	2.01 (0.20)	0.20	94.8
PQL type	1.82 (0.41)	0.38	90.6	1.81 (0.28)	0.28	89.0	1.82 (0.21)	0.20	87.6
$p_v(h)$	2.06 (0.41)	0.41	94.2	2.03 (0.28)	0.27	94.4	2.01 (0.21)	0.20	94.8

†Standard deviations are given in parentheses.

effects in λ, the variance of random effects, without using the EQL; the h-likelihood provides satisfactory analysis without any indication of bias (Noh *et al.*, 2006). The EQL is useful when, for a given variance function, the marginal likelihood for the GLM family does not exist.

Our h-likelihood is a full likelihood, whereas the empirical Bayes (EB) procedure uses a composite likelihood, omitting components from the full likelihood (Varin and Vidoni, 2005), which uses the component $\log\{f_\theta(v|y)\}$ only from the full h-likelihood

$$h = \log\{f_\theta(v|y)\} + \log\{f_\theta(y)\}.$$

Because $f_\theta(v|y)$ involves the fixed parameters θ we should use the whole h-likelihood to reflect the uncertainty about θ; it is the other component $f_\theta(y)$ which carries the information about this. Thus the EB method cannot properly account for the uncertainty that is caused by estimating θ. Various complicated remedies have been suggested for the EB interval estimate (Carlin and Louis, 2000), but h-likelihood gives estimates of proper standard errors for both random and fixed parameter estimates (Lee and Nelder, 2005a).

Invariance

The likelihood principle of Birnbaum (1962) states that the marginal likelihood carries all the (relevant experimental) information in the data about the fixed parameters θ, so it should be used for inferences about θ. The extended likelihood principle of Bjørnstad (1996) is that the joint likelihood of the form $\log\{f_\theta(y, v)\}$ carries all the information in the data about the unobserved quantities v and θ. However, this principle does not suggest how to use this likelihood for statistical analysis. It tells us only that some joint likelihod should serve as the basis for such an analysis. Given that some joint likelihood should serve as the basis for statistical inferences of HGLMs, we want to find a particular one whose maximization gives meaningful estimators of the random parameters. We have shown how maintaining invariance of inferences from the joint likelihood for trivial re-expressions of the underlying model leads to a unique definition of the h-likelihood, satisfying 'additivity on the linear predictor scale'. Lee and Nelder (2005a) gave a detailed discussion on how the h-likelihood gives a proper extension of the Fisher likelihood to random-effect models.

There have been several attempts to make likelihood inferences by using the joint likelihood. However, these previous attempts have not been successful, except Lee and Nelder's (1996). Consider the example of Bayarri *et al.* (1988):

$$y|u \sim \exp(u) \quad \text{and} \quad u \sim \exp(\theta).$$

This is Firth's first example when $\theta = 1$. The maximum likelihood estimator for random effects depends on the scale of u. So Bayarri *et al.* (1988) and Firth both conclude that likelihood inferences are not possible

for unobservables. Suppose that

$$h_1 = \log\{f_\theta(y,u)\} \text{ and } h_2 = \log\{f_\theta(y,v)\},$$

where $v = -\log(u)$. Here both $p_u(h_1)$ and $p_v(h_2)$ give the maximum likelihood estimator $\hat{\theta} = y$. Use of h_1 gives

$$1/\tilde{u} = \widehat{E(1/u|y)} = \hat{\theta} + y,$$

whereas h_2 gives

$$\tilde{u} = \widehat{E(u|y)} = 2/(\hat{\theta} + y);$$

both give empirical best predictors for different scales $E(1/u|y)$ and $E(u|y)$ respectively. In this example with $\theta = 1$ there is no obvious scale for the random effects to achieve additivity because there is neither a fixed nor an additional random effect. With the second example he shows that that 'additivity on the linear predictor scale' can lead to two different scales; for example it leads to the log-link and the identity link for random parameters. Lee and Nelder (2005a) gave an alternative way of determining a canonical scale for the h-likelihood. When the choice is not obvious we recommend the scale which makes the range of v to be the whole real line. Here, we define the h-likelihood with $v = \log(u)$ and find that $p_v(h)$ gives numerically satisfactory statistical inference (unreported). Further studies are necessary for general models where the choice of scale in random effects is not clear. In summary, in both of Firth's examples $p_v(h)$ gives satisfactory estimates for fixed parameters.

Note that the original adjusted profile likelihood of Barndorff-Nielsen (1983), to eliminate nuisance fixed effects β

$$p_\beta(m) = \left(m - \log\left[\frac{\det\{D(m,\beta)/2\pi\}}{2} \right] \right) \Big|_{\beta=\hat{\beta}} + \log\left\{ \det\left(\frac{\partial\hat{\beta}}{\partial\hat{\beta}_\tau} \right) \right\},$$

where $\tau = (\phi, \lambda)$, also includes an intractable Jacobian term and it is the cunning parameterization, allowing parameter orthogonality $E(\partial^2 m/\partial\beta\,\partial\tau) = 0$ of Cox and Reid (1987), which makes the Jacobian term $\log\{\det(\partial\hat{\beta}/\partial\hat{\beta}_\tau)\} \approx 0$. Similarly, the h-likelihood procedure overcomes a difficulty that is associated with Jacobian terms by choosing a proper scale for random effects. In $p_\beta(m)$ if orthogonal parameterization between fixed parameters does not hold the Jacobian term may not be ignorable, which limits the use of $p_\beta(m)$. However, for models without an apparent parameterization of random parameters $p_v(h)$ still gives satisfactory estimation for fixed parameters as indicated above.

Conclusion

Marginal likelihood has been considered as a basis of inferences for fixed parameters, complemented with EB estimation for random effects. Numerical integration, MCMC sampling etc. are often computationally too intensive (and so not feasible) and other methods such as PQL and marginal quasi-likelihood are severely biased.

Therefore, our h-likelihood procedure provides a valuable third way in between marginal and Bayesian methodology. In our view our approach completes Fisher likelihood by allowing inferences for random parameters without resorting to EB methods.

Finally, we accept the need for much theoretical work on DHGLMs, which we have not attempted to provide. There is scope for simulation to compare the properties of estimates by using different methods, and we hope that others will contribute both data sets and results of such simulations.

References in the discussion

Aalen, O. O. (1988) Heterogeneity in survival analysis. *Statist. Med.*, **7**, 1121–1137.

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

Barndorff-Nielsen, O. E. and Shephard, N. (2001) Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics (with discussion). *J. R. Statist. Soc.* B, **63**, 167–241.

Bayarri, M. J., DeGroot, M. H. and Kadane, J. B. (1988) What is the likelihood function (with discussion)? In *Statistical Decision Theory and Related Topics IV*, vol. 1 (eds S. S. Gupta and J. O. Berger). New York: Springer.

Bellio, R. and Varin, C. (2005) A pairwise likelihood approach to generalized linear models with crossed random effects. *Statist. Modllng*, **5**, 217–227.
Birnbaum, A. (1962) On the foundations of statistical inference (with discussion). *J. Am. Statist. Ass.*, **57**, 269–306.
Bjørnstad, J. F. (1996) On the generalization of the likelihood function and likelihood principle. *J. Am. Statist. Ass.*, **91**, 791–806.
Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear mixed models. *J. Am. Statist. Ass.*, **88**, 9–25.
Breslow, N. E. and Lin, X. (1995) Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, **82**, 81–91.
Browne, W. J. (1998) Applying MCMC methods to multi-level models. *PhD Thesis*. University of Bath, Bath.
Browne, W. J., Draper, D., Goldstein, H. and Rasbash, J. (2002) Bayesian and likelihood methods for fitting multilevel models with complex level-1 variation. *Computnl Statist. Data Anal.*, **39**, 203–225.
Carlin, B. P. and Louis, T. A. (2000) *Bayesian and Empirical Bayesian Methods for Data Analysis*. London: Chapman and Hall.
Chen, K., Hu, I. and Ying, Z. (1999) Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *Ann. Statist.*, **27**, 1155–1163.
Cox, D. R. and Reid, N. (1987) Parameter orthogonality and approximate conditional inference. *J. R. Statist. Soc.* B, **49**, 1–18.
Cox, D. R. and Reid, N. (2004) A note on pseudo-likelihood constructed from marginal densities. *Biometrika*, **91**, 729–737.
Duffie, D. and Singleton, K. (1993) Simulated moments estimation of Markov models of asset prices. *Econometrica*, **61**, 929–952.
Eberlein, E. and Keller, U. (1995) Hyperbolic distributions in finance. *Bernoulli*, **1**, 281–299.
Efron, B. (1986) Double exponential families and their use in generalized linear regression. *J. Am. Statist. Ass.*, **81**, 709–721.
Gallant, A. and Tauchen, G. (1996) Which moments to match. *Econometr. Theory*, **12**, 657–681.
Goldstein, H. (1986) Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika*, **73**, 43–56.
Goldstein, H. and Rasbash, J. (1996) Improved approximations for multilevel models with binary responses. *J. R. Statist. Soc.* A, **159**, 505–513.
Gourieroux, C., Monfort, A. and Renault, E. (1993) Indirect inference. *J. Appl. Econometr.*, **8**, S85–S118.
Ha, I. D. and Lee, Y. (2005) Comparison of hierarchical likelihood versus orthodox BLUP approach for frailty models. *Biometrika*, **92**, 717–723.
Hougaard, P. (2000) *Analysis of Multivariate Survival Data*. New York: Springer.
Lane, P. W. and Nelder, J. A. (1982) Analysis of covariance and standardization as instances of prediction. *Biometrics*, **38**, 613–621.
Lee, Y. (2004) Estimating intraclass correlation for binary data using extended quasi-likelihood. *Statist. Modllng*, **4**, 113–126.
Lee, Y. and Nelder, J. A. (1996) Hierarchical generalized linear models (with discussion). *J. R. Statist. Soc.* B, **58**, 619–678.
Lee, Y. and Nelder, J. A. (2000) The relationship between double-exponential families and extended quasi-likelihood families, with application to modelling Geissler's human sex ratio data. *Appl. Statist.*, **49**, 413–419.
Lee, Y. and Nelder, J. A. (2001a) Hierarchical generalised linear models: a synthesis of generalised linear models, random effect models and structured dispersions. *Biometrika*, **88**, 987–1006.
Lee, Y. and Nelder, J. A. (2001b) Modelling and analysing correlated non-normal data. *Statist. Modllng*, **1**, 3–16.
Lee, Y. and Nelder, J. A. (2005a) Likelihood for random-effect models (with discussion). *Statist. Oper. Res. Trans.*, **29**, 141–182.
Lee, Y. and Nelder, J. A. (2005b) Fitting via alternative random-effect models. *Statist. Comput.*, to be published.
Leppik, I. E., Dreifuss, F. E., Porter, R., Bowman, T., Santilli, N., Jacobs, M., Crosby, C., Cloyd, J., Stackman, J. and Graves, N. (1987) A controlled study of progabide in partial seizures: methodology and results. *Neurology*, **37**, 963–968.
Liang, K.-Y. and Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.
Lin, X. and Breslow, N. E. (1996) Bias correction in generalized linear mixed models with multiple components of dispersion. *J. Am. Statist. Ass.*, **91**, 1007–1016.
Lindsay, B. G. (1988) Composite likelihood methods. *Contemp. Math.*, **80**, 221–239.
Ma, R. (1999) An orthodox BLUP approach to generalized linear mixed models. *PhD Thesis*. University of British Columbia, Vancouver.
MacKenzie, G., Ha, I. D. and Lee, Y. (2006) Multivariate survival models based on the GTDL. Submitted to *Biostatistics*.
MacKenzie, G. and Pan, J. X. (2006) Optimal joint-mean covariance modelling. In *Select. Proc. 2nd Int. Wrkshp Correlated Data Modelling* (eds D. Gregori, G. MacKenzie, H. Friedl and R. Corradetti). Milan: Agneli. To be published.

Madan, D. B. and Seneta, E. (1990) The variance gamma model for share market returns. *J. Bus.*, **63**, 511–524.
Molenberghs, G. and Verbeke, G. (2005) *Models for Discrete Longitudinal Data*. New York: Springer.
Nelder, J. A. (1965) The analysis of randomized experiments with orthogonal block structure: I, Block structure and the null analysis of variance; II, Treatment structure and the general analysis of variance. *Proc. R. Soc. Lond.* A, **283**, 147–178.
Nelder, J. A. and Lee, Y. (1991) Generalized linear models for the analysis of Taguchi-type experiments. *Appl. Stochast. Mod. Data Anal.*, **7**, 107–120.
Nelder, J. A. and Wedderburn, R. W. M. (1972) Generalized linear models. *J. R. Statist. Soc.* A, **135**, 370–384.
Noh, M., Ha, I. D. and Lee, Y. (2006) Dispersion frailty models and HGLMs. *Statist. Med.*, to be published.
Noh, M. and Lee, Y. (2005) Robust modelling for inference from GLM classes. *Manuscript*. To be published.
Noh, M., Lee, Y. and Pawitan, Y. (2005) Robust ascertainment-adjusted parameter estimation. *Genet. Epidem.*, **29**, 68–75.
Pan, J. X. and MacKenzie, G. (2003) On modelling mean-covariance structures in longitudinal studies. *Biometrika*, **90**, 239–244.
Payne, R. W. (2004) Confidence intervals and tests for contrasts between combined effects in generally balanced designs. In *COMPSTAT 2004: Proc. Computational Statistics*, pp. 1629–1636. Heidelberg: Physica.
Payne, R. W. and Tobias, R. D. (1992) General balance, combination of information and the analysis of covariance. *Scand. J. Statist.*, **19**, 3–23.
Pourahmadi, M. (2000) Maximum likelihood estimation of generalized linear models for multivariate normal covariate matrix. *Biometrika*, **87**, 425–435.
Rao, C. R. (1965) The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika*, **52**, 355–372.
Rasbash, J., Browne, W. J., Healy, M., Cameron, B. and Charlton, C. (2004) *MLwiN (Version 2.0)*. London: Institute of Education.
Ridout, M. S., Demétrio, C. G. B. and Firth, D. (1999) Estimating intraclass correlation for binary data. *Biometrics*, **55**, 137–148.
Rigby, R. A. and Stasinopoulos, D. M. (2005) Generalized additive models for location, scale and shape (with discussion). *Appl. Statist.*, **54**, 507–554.
Rodríguez, G. and Goldman, N. (1995) An assessment of estimation procedures for multilevel models with binary responses. *J. R. Statist. Soc.* A, **158**, 73–89.
Sandmann, G. and Koopman, S. (1998) Estimation of stochastic volatility models via Monte Carlo maximum likelihood. *J. Econometr.*, **87**, 271–301.
Senn, S. (2000) Consensus and controversy in pharmaceutical statistics (with discussion). *Statistician*, **49**, 135–176.
Smyth, G. K. and Verbyla, A. P. (1999) Adjusted likelihood methods for modelling dispersion in generalized linear models. *Environmetrics*, **10**, 696–709.
Thall, P. F. and Vail, S. C. (1990) Some covariance-models for longitudinal count data with overdispersion. *Biometrics*, **46**, 657–671.
Varin, C. and Vidoni, P. (2005) A note on composite likelihood inference and model selection. *Biometrika*, **92**, 519–529.
Wedderburn, R. W. M. (1974) Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika*, **61**, 439–447.
Wolfinger, R. and O'Connell, M. (1993) Generalized linear mixed models: a pseudo-likelihood approach. *J. Statist. Computn Simuln*, **48**, 233–243.
Xia, N. and Zhang, Z. Z. (2005a) Convergence rate of Lee-Nelder estimators in Poisson-Gamma models. *Acta Math. Appl. Sin.*, to be published.
Xia, N. and Zhang, Z. Z. (2005b) Asymptotic normality of Lee-Nelder estimators in a HGLM family. *Working Paper*. Beijing University of Technology, Beijing.
Xue, X. (2001) Analysis of childhood brain tumour data in New York City using frailty models. *Statist. Med.*, **20**, 3459–3473.
Yates, F. and Cochran, W. G. (1938) The analysis of groups of experiments. *J. Agric. Sci.*, **28**, 556–580.
Yau, K. K. W. and Kuk, A. Y. C. (2002) Robust estimation in generalized linear mixed models. *J. R. Statist. Soc.* B, **64**, 101–117.