SORT 29 (2) July-December 2005, 141-164

Likelihood for random-effect models

Youngjo Lee[1] and John A. Nelder[2]

[1]*Seoul National University,* [2]*Imperial College London, U.K.*

Abstract

For inferences from random-effect models Lee and Nelder (1996) proposed to use hierarchical likelihood (h-likelihood). It allows inference from models that may include both fixed and random parameters. Because of the presence of unobserved random variables h-likelihood is not a likelihood in the Fisherian sense. The Fisher likelihood framework has advantages such as generality of application, statistical and computational efficiency. We introduce an extended likelihood framework and discuss why it is a proper extension, maintaining the advantages of the original likelihood framework. The new framework allows likelihood inferences to be drawn for a much wider class of models.

MSC: 62F10 62F15 62F30

Keywords: generalized linear models, hierarchical models, h-likelihood.

1 Introduction

Ever since Fisher introduced the concept of likelihood in 1921, the likelihood function has played an important part in the development of both the theory and the practice of statistics. The likelihood framework has advantages such as generality of application, algorithmic *wiseness* (Efron, 2003), consistency and asymptotic efficiency, which can be summarized as computational and statistical efficiency. Savage (1976) states "The most fruitful, and for Fisher, the usual definition of the likelihood associated with an observation is the probability or density of observation as a function of the parameter, modulo a multiplicative constant." Edwards (1972, pp. 12) similarly defines "the likelihood $L(H|R)$ of the hypothesis H given data R, and a specific model, to

Address for correspondence: Y. Lee, Department of Statistics, Seoul National University, San 56-1, Sillim-dong, Gwanak-gu, Seoul 151-742, Korea. E-mail: youngjo@plaza.snu.ac.kr
Received: October 2005

be proportional to $P(R|H)$, the constant of proportionality being arbitrary. However, when the problem does not fit into the usual parametric framework, the definition of the likelihood function is not immediately obvious.

Suppose that a model class consists of three types of object, observable random variables (data), unobservable (or unobserved) random variables and unknown fixed parameters. Special cases are subject-specific inferences for random-effect models, prediction of unobserved future observations, missing data problems etc. Consider the simplest example of a 2-level hierarchy with the model

$$y_{ij} = \beta + u_i + e_{ij},$$

where $u_i \backsim N(0, \lambda)$ and $e_{ij} \backsim N(0, \phi)$ with u_i and e_{ij} uncorrelated. This model leads to a specific multivariate distribution. From one point of view the parameters of the model are β, λ and ϕ, and it is straightforward to write down the likelihood from the multivariate normal distribution and to obtain estimates by maximizing it. However, although the u_i are thought of initially as having been obtained by sampling from a population, once a particular sample has been obtained they are fixed quantities and estimates of them will often be of interest (Searle et al 1992). The likelihood based upon the multivariate normal distribution provides no information on these quantities.

There have been several attempts to extend likelihood beyond its use in parametric inference to more general models that include unobserved random variables; see, for example, Henderson (1975) and Lee and Nelder (1996) for random-effect models, and Yates (1933) and Box *et al.* (1970) for missing data problems. Except for Lee and Nelder, these extensions have been successful only for inference of location parameters in limited classes of models. Pearson (1920) pointed out the limitation of Fisher likelihood inferences in prediction. As a likelihood solution various predictive likelihoods have been proposed (Bjørnstad, 1990): see Barndorff-Nielsen and Cox (1996) for interval estimates. Interpretation of the profile predictive likelihood approach of Mathiasen (1979) in the h-likelihood perspective is in Pawitan (2001, Chapter 16). In this paper we concentrate on likelihood inferences for random effects.

Bayarri *et al.* (1988) considered the following example: There is a single fixed parameter θ, a single unobervable random quantity U and a single observable quantity Y. The unobserved random variable U has a probability function

$$f_\theta(u) = \theta \exp(-\theta u) \quad \text{for } u > 0, \quad \theta > 0,$$

and an observable random variable Y has conditional probability function

$$f_\theta(y|u) = f(y|u) = u \exp(-uy) \quad \text{for } y > 0, \quad u > 0,$$

free of θ. Throughout this paper we use $f_\theta()$ as probability functions of random variables with fixed parameters θ; the arguments within the brackets can be either conditional or

unconditional. Thus, $f_\theta(y|u)$ and $f_\theta(u|y)$ have different functional forms even though we use the same $f_\theta()$ to mean probability functions with parameters θ.

Starting from a basic definition that the likelihood function is proportional to $L(r|\theta)$ where r and θ denote two derived classes, Bayarri *et al.* (1988) argue that there is no unique way of deciding which of the three classes should be regarded as part of r and which part of θ and, furthermore, that there is no unique way of deciding which random variables and parameters should explicitly enter the likelihood function. They consider three possibilities for an extended-likelihood for three objects:

$$\begin{aligned} L(y|\theta) &\equiv f_\theta(y) = \int f(y|u) f_\theta(u) du = \theta/(\theta + y)^2, \\ L(y|u, \theta) &\equiv f(y|u) = u \exp(-uy), \\ L(u; y|\theta) &\equiv f(y|u) f_\theta(u) = u\theta \exp\{-u(\theta + y)\}. \end{aligned}$$

Here "$L(y|\theta) \equiv f_\theta(y)$" means that $L(y|\theta) = f_\theta(y)c(y)$ for some $c(y) > 0$. The *marginal* likelihood $L(y|\theta)$ gives the (marginal) maximum-likelihood (ML) estimator $\hat{\theta} = y$, but is totally uninformative about the unknown value of u of U. The *conditional* likelihood in the form $L(y|u, \theta)$, which may be regarded as

$$L(\text{observed}|\text{unobserved}),$$

is uninformative about θ and loses the relationship between u and θ reflected in $f_\theta(u)$. Finally, the *joint* likelihood $L(u; y|\theta)$, which may be regarded as

$$L(\text{random variables}|\text{parameters}),$$

yields, if maximized with respect to θ and u, the useless estimators $\hat{\theta} = \infty$ and $\hat{u} = 0$. Bayarri *et al.* (1988) therefore concluded that none is useful as a likelihood for more general inferences.

In extended likelihood dividing the three types of object into two derived classes would be confusing. For example the empirical Bayes method uses $f_\theta(u|y)$, which seems to belong to $L(\text{observed}|\text{unobserved})$. If so it cannot be distinguished from $L(y|u, \theta) \equiv f_\theta(y|u)$. In this paper $L(a; b)$ denotes the likelihood for the argument a using the probability function $f_\theta(b)$. Here $L(\theta, u; u|y) \equiv f_\theta(u|y)$ and $L(\theta, u; y|u) \equiv f_\theta(y|u)$. We use capital letters such as L for likelihood and lowercase letters such as $l = \log L$ for log likelihood which we shall abbreviate to *loglihood* (a useful contraction which we owe to Michael Healy).

In this paper we resolve Bayarri *et al.*'s (1988) problem by showing that it is possible to make unambiguous likelihood inferences about a wide class of models having unobserved random variables. If in this example, instead of the joint likelihood $L(\theta, u; y, u)$, we use a particular form of it, the so-called h-likelihood

$$L(\theta, u; y, \log u) \equiv f(y|\log u) f_\theta(\log u) = u^2 \theta \exp\{-u(\theta + y)\}$$

maximization gives the ML estimator $\hat{\theta} = y$, and the random effect estimator $\hat{u} = 1/y$.

When θ is known, the best predictor (BP; Searle et al 1992, pp. 261) of u is defined by

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

When θ is unknown, the h-likelihood gives the empirical BP of u

$$\hat{u} = \widehat{E(u|y)} = 2/(\theta + y)|_{\theta=\hat{\theta}} = 1/y.$$

The word *predictor* has often been used for random-effect estimates. For the prediction of unobserved future observations we believe that it is the right one to use. However, for inference about unknown random effects the word *estimate* seems more appropriate because we are estimating unknown u, fixed once the data y are known, though possibly changing in future samples.

Our goal is to establish an extended likelihood framework by showing that the h-likelihood is a proper extension of the Fisher likelihood to random-effect models, maintaining the original likelihood framework for parametric inferences. In Section 2 we define the h-likelihood for hierarchical generalized linear models (HGLMs). In Section 3, we show why the h-likelihood, among joint likelihoods, should be used, and in Section 4 we illustrate why we need to distinguish two classes of parameters, fixed effects (for location) and dispersions. In Section 5, we describe the extended likelihood framework, and in Section 6 we illustrate extended likelihood inferences using Bayarri *et al.*'s (1988) example. We also explain how our method differs from Breslow and Clayton's (1993) penalized-quasi-likelihood (PQL) method and illustrate why the latter suffers from severe bias. In Section 7 we discuss how the extended framework preserves the advantages of the original likelihood framework, giving our conclusions in Section 8.

2 HGLMS

HGLMs are generalized linear models (GLMs) in which the linear predictor contains both fixed and random parameters: They take the form

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

with a linear predictor

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

where $g()$ is a GLM link function, X and Z are model matrices for fixed and random parameters (effects) respectively, and $v_i = v(u_i)$ are random effects after some transformation $v()$. For simplicity we consider the case of just one random vector u.

Here the joint density of the responses y and the random effects u can be used to define a *joint likelihood*

$$L(\theta, u; y, u) \equiv f_{\beta,\phi}(y|u) f_\lambda(u), \tag{2}$$

where $\theta = (\beta, \phi, \lambda)$. In (2) $f_{\beta,\phi}(y|u)$ is a density with a distribution from a one-parameter exponential family, while the second term $f_\lambda(u)$ is the density function of the random effects u with parameter λ.

2.1 H-likelihoods

We call $L(\theta, u; y, u)$ a joint likelihood, a phrase first used by Henderson (1975) in the context of linear mixed models; in our notation these are normal-normal HGLMs, in which the first element refers to the distribution of $y|u$ and the second to that for u. A joint likelihood is not a likelihood in the Fisherian sense because of the presence of unobservables, namely the random effects. Bjørnstad (1996) showed that the joint likelihood satisfies the likelihood principle that the likelihood of the form $L(\theta, u; y, u)$ carries all the (relevant experimental) information in the data about the unobserved quantities u and θ (Edwards 1972, pp.30 and Berger 1985, pp.28). However, the likelihood principle does not provide any obvious suggestions on how to use this likelihood for statistical analysis. It tells us only that some joint likelihood should serve as the basis for such an analysis. For example, the use of $L(\theta, u; y, u)$ in Bayarri *et al.*'s (1988) example results in useless inferences about the random parameter.

A joint likelihood $L(\theta, u; y, k(u))$ is not in general invariant with respect to the choice of parametrization $k(u)$ of the random parameter u, because a change in this parametrization involves a Jacobian term for u. Lee and Nelder (1996) proposed to use the joint density of y and the random effects $v = v(u)$ on the particular scale as shown in (1) to form a subclass of joint likelihoods,

$$L(\theta, v; y, v) \equiv f_{\beta,\phi}(y|v) f_\lambda(v). \tag{3}$$

These were called h-likelihoods by Lee and Nelder (1996), who used them as extended likelihoods for HGLMs. Even though $f_{\beta,\phi}(y|v(u)) \equiv f_{\beta,\phi}(y|u)$ mathematically, we write the conditional density as $f_{\beta,\phi}(y|v(u))$ to stress that the function $v(u)$ defines the scale on which the random effects are assumed to combine additively with the fixed effects β in the linear predictor.

3 Why h-likelihoods among joint likelihoods?

Given that some joint likelihood should serve as the basis for statistical inferences of a general nature, we want find a particular one whose maximization gives meaningful estimators of the random parameters. 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. For further development we need the following property of joint likelihoods.

Property. The joint likelihoods $L(\theta, u; y, u)$ and $L(\theta, u; y, k(u))$ give identical inferences about the random effects if $k(u)$ is a linear parametrization of u.

This property of joint likelihoods is meaningful because the BP property can be preserved only under linear transformation, i.e. $E\{k(u)|y\} = k(E(u|y))$ only if $k()$ is linear.

Consider a simple normal-normal HGLM of the form: for $i = 1, \ldots, m$ and $j = 1, \ldots, n$ with $N = mn$

$$y_{ij} = \beta + v_i + e_{ij}, \tag{4}$$

where $v_i \backsim i.i.d.\ N(0, \lambda)$ and $e_{ij} \backsim i.i.d.\ N(0, 1)$. Consider a linear transformation $v_i = \sigma v_i^*$ where $\sigma = \lambda^{1/2}$ and $v_i^* \backsim i.i.d.\ N(0, 1)$. The joint loglihoods $l(\theta, v; y, v)$ and $l(\theta, v^*; y, v^*)$ give the same inference for v_i and v_i^*. In (3) the first term $\log f_{\beta,\phi}(y|v)$ is invariant with respect to reparametrizations; in fact $f_{\beta,\phi}(y|v) = f_{\beta,\phi}(y|u)$ functionally for one-to-one parametrization $v = v(u)$. Let $\hat{v}_i$ and $\hat{v}_i^*$ maximize $l(\theta, v; y; , v)$ and $l(\theta, v^*; y, v^*)$, respectively. Then, we have invariant estimates $\hat{v}_i = \sigma \hat{v}_i^*$ because

$$-2 \log f_\lambda(v) = m \log(2\pi\sigma^2) + \sum v_i^2/\sigma^2 = -2 \log f_\lambda(v^*) + m \log(\sigma^2),$$

these loglihoods differ only by a constant.

Consider now model (4), but with a parametrization

$$y_{ij} = \beta + \log u_i + e_{ij}, \tag{5}$$

where $\log(u_i) \backsim i.i.d.\ N(0, \lambda)$. Let $\log(u_i) = \sigma \log u_i^*$ and $\log(u_i^*) \backsim i.i.d.\ N(0, 1)$. Here we have

$$\begin{aligned} -2 \log f_\lambda(u) &= m \log(2\pi\lambda) + \sum (\log u_i)^2/\lambda + 2 \sum \log u_i \\ &= -2 \log f_\lambda(u^*) + m \log(\lambda) + 2 \sum \log(u_i/u_i^*). \end{aligned}$$

Let $\hat{u}_i$ and $\hat{u}_i^*$ maximize $l(\theta, u; y; , u)$ and $l(\theta, u^*; y, u^*)$, respectively. Then, $\log \hat{u}_i \neq \sigma \log \hat{u}_i^*$ because $\log u_i = \sigma \log u_i^*$, i.e. $u_i = u_i^{*\sigma}$, is no longer a linear transformation.

Clearly the two models (4) and (5) are equivalent, so that if h-likelihood is to be a

useful notion we need their corresponding h-loglihoods be equivalent as well. In fact the h-likelihood for model (5) is

$$L(\theta, v; y, v) \equiv f_{\beta,\phi}(y|\log u) f_\lambda(\log u)$$

in accordance with the rule that the random effect appears linearly in the linear predictor on the scale $v = \log u$, giving

$$\eta_{ij} = \mu_{ij} = \beta + v_i \quad \text{with} \quad \mu_{ij} = E(y_{ij}|v_i).$$

To maintain invariance of inference with respect to equivalent modellings, we must define the h-likelihood on the particular scale $v(u)$ on which the random effects combine additively with the fixed effects β in the linear predictor.

For simplicity of argument, let $\lambda = 1$, so that there is no dispersion parameter, but only a location parameter β. The h-loglihood $l(\theta, v; y, v)$ is given by

$$-2h = -2l(\theta, v; y, v) \equiv \{N \log(2\pi) + \sum_{ij}(y_{ij} - \beta - v_i)^2\} + \{m \log(2\pi) + \sum_i v_i^2\}.$$

This has its maximum at the BP

$$\hat{v}_i = E(v_i|y) = \frac{n}{n+1}(\bar{y}_{i.} - \beta).$$

Suppose that we estimate β and v by joint maximization of h. The solution is

$$\hat{\beta} = \bar{y}_{..} = \sum_{ij} y_{ij}/N \quad \text{and} \quad \hat{v}_i = \frac{n}{n+1}(\bar{y}_{i.} - \bar{y}_{..}) = \sum_j (y_{ij} - \bar{y}_{..})/(n+1).$$

Now $\hat{\beta}$ is the ML estimator and $\hat{v}_i$ is the empirical BP defined by

$$\hat{v}_i = \widehat{E(v_i|y)},$$

and can be also justified as the best linear unbiased predictor (BLUP; Searle *et al.* 1992, pp. 269).

The joint loglihood $L(\beta, u; y, u)$ gives

$$-2L(\beta, u; y, u) \equiv \{N \log(2\pi) + \sum (y_{ij} - \beta - \log u_i)^2\} + \{m \log(2\pi) + \sum (\log u_i)^2 + 2 \sum (\log u_i)\} \tag{6}$$

with an estimate

$$\hat{v}_i = \log \hat{u}_i = \frac{n}{n+1}(\bar{y}_{i.} - \beta) - 1/(n+1).$$

The joint maximization of $L(\beta, u; y, u)$ leads to

$$\hat{\beta} = \bar{y}_{..} + 1 \text{ and } \hat{v}_i = \frac{n}{n+1}(\bar{y}_{i.} - \bar{y}_{..}) - 1.$$

Thus, in this example joint maximization of the h-loglihood provides satisfactory estimates of both the location and random parameters for either parameterization, while that of a joint loglihood may not.

4 Is joint maximization valid for estimation of dispersion parameters?

Lee and Nelder (1996, 2001*a*) distinguished two types of parameters, fixed effects (location parameters) β and dispersion parameters (ϕ, λ). The use of restricted maximum likelihood (REML) shows that different functions must be maximized to estimate location and dispersion parameters. Our generalization of REML shows that an appropriate adjusted profile h-likelihood (APHL) should be used for estimation of dispersion parameters (Lee and Nelder, 1996, 2001).

Consider the following two equivalent non-normal models: for $i = 1, \ldots, m$

$$y_i|u_i \backsim Poisson(\delta u_i) \quad \text{and} \quad u_i \backsim \exp(1), \tag{7}$$

and

$$y_i|w_i \backsim Poisson(w_i) \quad \text{and} \quad w_i \backsim \exp(1/\delta), \tag{8}$$

where $w_i = \delta u_i$; so we have $E(u_i) = 1$ and $E(w_i) = \delta$.

Note that while fixed effects β appear only in $f_{\beta,\phi}(y|v)$, dispersion parameters (ϕ, λ) can appear in both $f_{\beta,\phi}(y|v)$ and $f_\lambda(v)$. In model (7), use of the log link, on which the fixed and random effects are additive, leads to

$$\log \mu_i = \beta + v_i,$$

where $\mu_i = E(y_i|u_i) = E(y_i|w_i)$, $\beta = \log \delta$, and $v_i = \log u_i$. Now β is a fixed effect, so that $\delta = \exp(\beta)$ is a location parameter in the HGLM context.

In model (8) there is only one random component and no fixed effect, so that the choice of link function, and therefore of $v(u)$, is arbitrary. With an identity link, δ is no longer a fixed effect but becomes the dispersion parameter λ appearing in $f_\lambda(w)$, for which the maximized h-likelihood maintains invariance only with respect to translations. Lee and Nelder (1996, 2001*a*) proposed a different estimation scheme for such parameters as we shall see in the next Section.

We now show that the joint maximization of the h-loglihoods cannot be used for estimation of the dispersion parameters. Suppose that we have an identity link in model

(8). Then h-loglihoods are $L(\theta, u; y, u)$ and $L(\theta, w; y, w)$ for the linear transformation $w = \delta u$. Then,

$$\log f(u) = -\sum u_i = -\sum w_i/\delta = \log f_\delta(w) + m \log \delta$$

so that given δ, random-effect predictions are given by $\hat{u}_i = \hat{w}_i/\delta = y_i/(\delta + 1)$. However, for δ the maximization of $L(\theta, u; y, u)$ yields an estimating equation $\sum y_i = \delta \sum \hat{u}_i = \delta \sum y_i/(\delta + 1)$ with a solution $\hat{\delta} = \infty$ and the use of $L(\theta, w; y, w)$ yields an estimating equation $\delta = \sum \hat{w}_i/m = \delta \sum y_i/\{m(\delta + 1)\}$ with a solution $\hat{\delta} = \bar{y}_\cdot - 1$ where $\bar{y}_\cdot = \sum y_i/m$. Thus, different estimates for the dispersion parameter δ are obtained by jointly maximizing the h-loglihoods $L(\theta, u; y, u)$ and $L(\theta, w; y, w)$ from the same model (8) but with a different parametrization $w = \delta u$.

In model (7), use of the log link leads to the h-loglihoods $L(\theta, u; y, \log u)$ and $L(\theta, w; y, \log w)$ for linear transformation $z = \beta + v$ where $v = \log u$ and $z = \log w$. Here, we have

$$\log f(v) = \sum(-u_i + v_i) = \sum(-w_i/\delta + z_i - \log \delta) = \log f_\delta(z). \tag{9}$$

The joint maximization of this h-loglihood $L(\theta, u; y, \log u)$ with respect to δ and v_i gives

$$\hat{u}_i = (y_i + 1)/(\hat{\delta} + 1) = \widehat{E(u_i|y_i)}$$

because $E(u_i|y_i) = (y_i + 1)/(\delta + 1)$, and the marginal ML estimator $\hat{\delta} = \bar{y}$. Similarly, joint maximization of $L(\theta, w; y, \log w)$ with respect to δ and z_i gives

$$\hat{w}_i = \hat{\delta}(y_i + 1)/(\hat{\delta} + 1) = \widehat{E(w_i|y_i)}$$

because $E(w_i|y_i) = \delta(y_i + 1)/(\delta + 1)$, so also $\hat{z}_i = \hat{v}_i + \log \hat{\delta}$, and again the marginal ML estimator is given by $\hat{\delta} = \bar{y}_\cdot$ Thus, identical estimates for the location parameter δ are obtained by jointly maximizing the h-loglihoods $L(\theta, u; y, \log u)$ and $L(\theta, w; y, \log w)$.

In multiplicative models such as (7) because

$$E(y_i|u_i) = \delta u_i = (c\delta)(u_i/c) \text{ for any } c > 0$$

we may put constraints on either the random effects or the fixed effects. Lee and Nelder (1996) proposed to put constraints on the random effects; for example in model (7) we put $E(u_i) = 1$ which is convenient when there is more than one random component. This strategy converts model (8) to an equivalent model (7), where the log link is an obvious choice in forming the h-loglihood, giving invariant inference for the fixed effect δ. Thus, putting constraints on random effects enlarges the set of fixed effects $\beta = \log \delta$ which can be estimated by a direct maximization of h. We recommend following this strategy in defining h-loglihoods, though it is not compulsory.

In the multiplicative model above neither u_i nor β is separately identifiable because they depend upon an arbitrary constraint. In an additive model such as (4) β is identifiable as $E(y_{ij})$ because of constraints $E(v_i) = 0$ and $E(e_{ij}) = 0$. The model (4) assumes $E(v_i) = 0$ and the model (7) assumes $E(u_i) = 1$, so that care is necessary in comparing parameter estimates from different models. Lee and Nelder (2004) showed that differences in the behaviour of parameters between random-effect models and GEE models are caused by assuming different constraints and therefore are based on a failure to compare like with like.

5 Extended likelihood framework

The original Fisher likelihood framework had two types of object and two kinds of inference:
Data Generation: Generate an instance of the data y from a probability function with given fixed parameters θ

$$f_\theta(y).$$

Parameter Estimation: Given the data y, make an inference about an unknown fixed θ in the stochastic model by using the likelihood

$$L(\theta; y).$$

The connection between these two processes is given by

$$L(\theta; y) \equiv f_\theta(y),$$

where L and f are algebraically identical, but on the left-hand side y is fixed while θ varies and on the right-hand side θ is fixed while y varies.

The extended likelihood framework for three types of object can be described as follows:
Data Generation: (i) Generate an instance of the random effects v from a probability function $f_\theta(v)$ and then with v fixed, (ii) generate an instance of the data y from a probability function $f_\theta(y|v)$. The combined stochastic model is given by the product of the two probability functions

$$f_\theta(v) f_\theta(y|v). \tag{10}$$

Parameter Estimation: Given the data y, we can (i) make inferences about θ by using the marginal likelihood $L(\theta; y) \equiv f_\theta(y)$, and (ii) given θ, make inferences about v by using

the conditional likelihood in the form

$$L(\theta, v; v|y) \equiv f_\theta(v|y).$$

Given the data y, the extended likelihood for the joint unknowns (v, θ) is given by

$$L(\theta, v; y, v) = L(\theta; y)L(\theta, v; v|y) \equiv f_\theta(y)f_\theta(v|y). \tag{11}$$

The connection between these two processes is given by

$$f_\theta(y)f_\theta(v|y) \equiv L(\theta, v; y, v) \equiv f_\theta(v, y) = f_\theta(v)f_\theta(y|v). \tag{12}$$

On the left-hand side y is fixed while (v, θ) vary, while on the right-hand side θ is fixed while (v, y) vary. In the extended likelihood framework the v appear in data generation as random instances, but in parameter estimation as unknowns.

The combined stochastic model $f_\theta(y|v)f_\theta(v)$ in (10) for data generation is often easily available in an explicit form. However, the practical difficulties in extended likelihood inference stem from the fact that the two components,

$$f_\theta(y) = \int f_\theta(y|v)f_\theta(v)dv \text{ and } f_\theta(v|y) = f_\theta(y|v)f_\theta(v)/\int f_\theta(y|v)f_\theta(v)dv,$$

are generally hard to obtain, except for some conjugate families, because of the integration involved. However, the h-likelihood can exploit the connection (12) to give likelihood inferences of a general nature. In the next Section we show how to implement inferential procedures without explicitly computing the two components $f_\theta(y)$ and $f_\theta(v|y)$.

Let

$$h = m + \log f_\theta(v|y) = \log f_\theta(v) + \log f_\theta(y|v), \tag{13}$$

where m is the marginal loglihood $m = \log L(\theta; y)$. This is the h-loglihood, which plays the same role as the loglihood in Fisher's likelihood inference.

5.1 Inference for random parameters

The relative simplicity of h-likelihood methods of inference becomes apparent when we compare them with other methods. If the conditional density $f_\theta(v|y)$ follows the normal distribution, it is immediate that given θ, the maximum h-likelihood estimator for v is a BP, i.e. $\hat{v} = E(v|y)$. If there exists a transformation $k()$ such that $L(\theta, v; v|y)$ $(\equiv f_\theta(k(v)|y))$ takes the form of a normal distribution, the h-likelihood gives a BP for $\widehat{k(v)} = k(\hat{v}) = E\{k(v)|y\}$. However, the h-likelihood gives more than this.

Consider a mixed linear model

$$Y = X\beta + Zv + e,$$

where $v \backsim MVN(0, \Lambda)$ and $e \backsim MVN(0, \Sigma)$ and MVN stands for a multivariate normal distribution. Henderson (1975) showed that the joint maximization of his joint loglihood (which, for the normal-normal model, is the h-loglihood) leads to the estimating equations

$$\begin{pmatrix} X^T\Sigma^{-1}X & X^T\Sigma^{-1}Z \\ Z^T\Sigma^{-1}X & Z^T\Sigma^{-1}Z + \Lambda^{-1} \end{pmatrix}\begin{pmatrix} \hat{\beta} \\ \hat{v} \end{pmatrix} = \begin{pmatrix} X^T\Sigma^{-1}Y \\ Z^T\Sigma^{-1}Y \end{pmatrix}.$$

Set H to be the square matrix of the left hand side, $V = Z\Lambda Z^T + \Sigma$, and $D = Z^T\Sigma^{-1}Z + \Lambda^{-1}$. Then, given Λ and Σ, the solution for β gives the ML estimator, satisfying

$$X^T V^{-1} X\hat{\beta} = X^T V^{-1} Y,$$

and the solution for v gives the empirical BPs

$$\hat{v} = \widehat{E(v|Y)} = E(v|Y)|_{\beta=\hat{\beta}} = \Lambda Z^T V^{-1}(Y - X\hat{\beta}) = D^{-1}Z^T\Sigma^{-1}(Y - X\hat{\beta}).$$

Furthermore, H^{-1} gives estimates of

$$E\left\{\begin{pmatrix} \hat{\beta} - \beta \\ \hat{v} - v \end{pmatrix}\begin{pmatrix} \hat{\beta} - \beta \\ \hat{v} - v \end{pmatrix}^T\right\}.$$

This yields $(X^T V^{-1} X)^{-1}$ as a variance estimate for β, which coincides with that for the ML estimates. Now we see that H^{-1} also gives the correct estimate for $E\left\{(\hat{v}-v)\,(\hat{v}-v)^t\right\}$.

When β is known we use the BP

$$\tilde{v} = E(v|Y).$$

Then we have

$$var(\tilde{v} - v) = E\left\{(\tilde{v} - v)(\tilde{v} - v)^T\right\} = E\left\{var(v|Y)\right\}.$$

Note here that

$$var(v|Y) = \Lambda - \Lambda Z^T V^{-1} Z\Lambda = D^{-1}.$$

When β is known D^{-1} gives a proper estimate of the variance of $\tilde{v} - v$.

The extended likelihood principle of Bjørnstad (1996) is that the joint likelihood of the form $L(\theta, v; y, v)$ carries all the information in the data about the unobserved

quantities v and θ. Because $f_\theta(y)$ in (11) does not involve v, $L(\theta, v; v|y) \equiv f_\theta(v|y)$ seems to carry all the information in the data about the random parameters. This leads to the empirical Bayes (EB) method for inference about v, which uses the estimated posterior

$$f_{\hat{\theta}}(v|y),$$

where $\hat{\theta}$ are usually the marginal ML estimators (Carlin and Louis, 2000). Thus, maximization of the h-likelihood yields EB-mode estimators, and they can be obtained without computing $f_\theta(v|y)$. However, when β is unknown the estimated posterior $f_{\hat{\beta}}(v|y)$ for the EB procedure gives $D^{-1}|_{\beta=\hat{\beta}}$ as a naive estimate for $var(\hat{v} - v)$, and this does not properly account for the uncertainty caused by estimating β. Various complicated remedies have been suggested for the EB interval estimate (Carlin and Louis, 2000).

By contrast, the h-loglihood gives a straightforward correction. Here we have

$$var(\hat{v} - v) = E\{var(v|Y)\} + E\left\{(\hat{v} - \tilde{v})(\hat{v} - \tilde{v})^T\right\},$$

where the second term shows the variance inflation caused by estimating the unknown β. As an estimate for $var(\hat{v} - v)$ the appropriate component of H^{-1} gives

$$\{D^{-1} + D^{-1}Z^T\Sigma^{-1}X(X^TV^{-1}X)^{-1}X^T\Sigma^{-1}ZD^{-1}\}|_{\beta=\hat{\beta}}.$$

Because $\hat{v} - \tilde{v} = -\Lambda Z^T V^{-1} X(\hat{\beta} - \beta)$ we can show that

$$E\left\{(\hat{v} - \tilde{v})(\hat{v} - \tilde{v})^T\right\} = D^{-1}Z^T\Sigma^{-1}X(X^TV^{-1}X)^{-1}X^T\Sigma^{-1}ZD^{-1}.$$

Thus, the h-loglihood handles correctly the variance inflation caused by estimating fixed effects. From this we can construct confidence bounds for unknown v, fixed once the data are observed. Lee and Nelder (1996) extended the results of this section to general HGLMs under some regularity conditions. Later, we illustrate this further by using Bayarri *et al.*'s (1988) example as a non-normal model.

We see that inferences about random effects cannot be made by using solely $f_\theta(v|y)$, as the EB method does. 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. The notation $L(\theta, v; v|y)$ shows that the EB problem is caused by the nuisance fixed parameters θ.

5.2 Inference for fixed parameters for both location and dispersion

The likelihood principle of Birnbaum (1962) is that the marginal likelihood $L(\theta; y)$ carries all the (relevant experimental) information in the data about the fixed parameters θ, so that $L(\theta; y)$ should be used for inferences about θ: see also Berger and Wolpert

(1984). For inferences about fixed parameters θ we can use $f_\theta(y)$ alone because $L(\theta; y)$ does not involve nuisance random parameters at all. However, in general the marginal likelihood requires intractable integration. One method of obtaining the marginal ML estimators for θ is the expectation-maximization (EM) algorithm of Dempster *et al.* (1977). This exploits the property (13) of joint loglihoods using the result that under appropriate regularity conditions

$$E(\partial h/\partial\theta|y) = \partial m/\partial\theta + E(\partial \log f_\theta(v|y)/\partial\theta|y) = \partial m/\partial\theta.$$

The last equality is immediate from the fact that

$$\int f_\theta(v|y)dv = 1.$$

The EM algorithm is often numerically slow to converge and it is analytically hard to evaluate the conditional expectation $E(h|y)$. Alternatively, simulation methods, such as Monte Carlo EM (Vaida and Meng, 2004) and Gibbs sampling (Karim and Zeger, 1992), can be used to evaluate the conditional expectation, but these methods are computationally intensive. Instead, numerical integration using Gauss-Hermite quadrature (Crouch and Spiegelman, 1990) could be directly applied to obtain the ML estimators, but this also becomes computationally heavier as the number of random components increases.

By contrast, we can obtain estimators for θ by directly maximizing appropriate quantities derived from the h-loglihood, and compute their standard error estimates from the second derivatives. In our framework we do not need to evaluate an analytically difficult expectation step nor use a computationally intensive method, such as Monte Carlo EM or numerical integration. Instead we maximize adjusted profile h-likelihoods (APHLs) to obtain ML and REML estimators. Ha and Lee (2005*a*) showed how h-likelihood gives straightforward estimators of β for mixed linear models with censoring, whereas the ordinary EM method has difficulty.

In our framework there are two useful adjusted profile loglihoods for inferences about fixed parameters. The marginal loglihood m can be obtained from the h-loglihood by integrating out the random parameters,

$$m \equiv \log \int f_\theta(v, y)dv = \log \int f_\theta(y|v)f_\theta(v)dv. \tag{14}$$

In mixed linear models the conditional density,

$$f_{\phi,\lambda}(y|\tilde{\beta}) = f_{\beta,\theta}(y)/f_{\beta,\theta}(\tilde{\beta}),$$

where $\tilde{\beta}$ are ML estimators given (ϕ, λ), is free of β (Smyth , 2002), so that the restricted loglihood of Patterson and Thompson (1971) can be written as

$$r = \log L(\phi, \lambda; y|\tilde{\beta}) \equiv \log f_{\phi,\lambda}(y|\tilde{\beta}).$$

This has been proposed for inference about the dispersion parameters (ϕ, λ) to reduce bias, especially in finite samples: see also Harville (1977).

In our framework the marginal loglihood m is an adjusted profile loglihood for the fixed parameters θ, after eliminating random parameters v by integration from the h-loglihood h, and the restricted loglihood r is that for the dispersion parameters (ϕ, λ), after eliminating fixed parameters β by conditioning from the marginal loglihood m. However, in general they hard to obtain because they use the marginal loglihood m. Let l be a loglihood, either a marginal loglihood m or an h-loglihood h, with nuisance parameter α, random or fixed. Lee and Nelder (2001a) considered a set of functions $p_\alpha(l)$, defined by

$$p_\alpha(l) = [l - \frac{1}{2}\log\det\{D(l,\alpha)/(2\pi)\}]|_{\alpha=\tilde{\alpha}} \tag{15}$$

where $D(l,\alpha) = -\partial^2 l/\partial\alpha^2$ and $\tilde{\alpha}$ solves $\partial l/\partial\alpha = 0$. For fixed effects β the use of $p_\beta(m)$ is equivalent to conditioning on $\tilde{\beta}$, i.e. $p_\beta(m) \simeq r = l(\phi, \lambda; y|\tilde{\beta}) \equiv \log f_{\phi,\lambda}(y|\tilde{\beta})$ to the first order (Cox and Reid, 1987), while for random effects v the use of $p_v(h)$ is equivalent to integrating them out using the first-order Laplace approximation, i.e. $p_v(h) \simeq m$ (Lee and Nelder 2001a). The set of functions $p_*()$ may be regarded as derived loglihoods for various subsets of parameters.

In mixed linear models

$$m \equiv p_v(h) \text{ and } p_\beta(m) \equiv p_{\beta,v}(h).$$

Thus, here the marginal ML estimators for β and their standard error estimates can be obtained by directly maximizing the adjusted profile h-loglihood $p_v(h)$, instead of the joint maximization of the previous Section. The restricted loglihood r has been used only in mixed linear models. Its natural extension is $p_\beta(m)$ (Cox and Reid, 1987). To avoid intractable integration, instead of $p_\beta(m)$ we may use $p_{\beta,v}(h)$ as the restricted likelihood for dispersion parameters. The use of $p_{\beta,v}(h)$ for estimating the dispersion parameters (ϕ, λ) means that we can eliminate both random and fixed effects simultaneously from the h-likelihood. Lee and Nelder (2001a) showed that in general $p_{\beta,v}(h)$ is approximately $p_\beta(p_v(h))$ and that numerically $p_{\beta,v}(h)$ provides good dispersion estimators for HGLMs. This reduces bias of the ML estimator greatly in frailty models with nonparametric baseline hazards where the number of nuisance β increases with sample (Ha and Lee, 2005b) .

In principle we should use the h-loglihood h for inferences about v, the marginal-loglihood m for β and the restricted loglihood $p_\beta(m)$ for the dispersion parameters.

When m is numerically hard to obtain, we propose to use APHLs $p_v(h)$ and $p_{\beta,v}(h)$ as approximations to m and $p_\beta(m)$; $p_{\beta,v}(h)$ gives approximate restricted ML estimators for the dispersion parameters and $p_v(h)$ approximate ML estimators for the location parameters. Higher-order approximations can be useful for improved accuracy (Lee and Nelder 2001*a*). Although in general a joint maximization of the h-loglihood does not provide marginal ML estimators for β the deviance differences constructed from h and $p_v(h)$ are often very similar, so that we propose to use h for estimating β unless it yields non-ignorable biases. For example, in HGLMs for binary data $p_v(h)$ should be used for estimating β (Noh and Lee 2004).

Using the formula

$$p_v(h) = [h - \frac{1}{2}\log\det\{D(h,v)/(2\pi)\}]|_{v=\tilde{v}},$$

model (4) gives $D(h,v) = diag(d_i)$ where $d_i = n + 1$, and model (7) gives $d_i = y_i + 1$, i.e. for both models $D(h,v)$ is independent of the fixed effects β, depending only upon dispersion parameters if these exist, so that the maximization of h also provides the ML estimators for β. This is true for three models, the normal-normal, Poisson-gamma and gamma-inverse gamma with log link; for these explicit forms for m are available (Lee and Nelder 1996).

Breslow and Clayton (1993) proposed the use of the REML estimating equations for normal mixed linear models to estimate dispersion parameters in GLMMs. In mixed linear models the adjustment term $D(h,\delta)$ with $\delta = (v,\beta)$ does not involve δ. However, this is not so in general, so that Breslow and Clayton's (1993) method suffers from severe bias because it ignores derivative terms $\partial\tilde{\delta}/\partial\lambda$ and $\partial\tilde{\delta}/\partial\phi$. Furthermore, Lin and Breslow's (1996) correction of the Breslow and Clayton's method still suffers from the non-ignorable bias caused by ignoring these important terms, while the h-likelihood procedure does not (Noh and Lee, 2004).

We now illustrate the h-likelihood approach for random-effect models using Bayarri *et al.*'s (1988) example.

6 Bayarri *et al.*'s example revisited

Let us return to Bayarri et al's (1988) example:

$$y|u \backsim \exp(u) \quad \text{and} \quad u \backsim \exp(\theta), \tag{16}$$

and equivalently

$$y|w \backsim \exp(w/\theta) \quad \text{and} \quad w \backsim \exp(1) \tag{17}$$

where $E(w) = 1$ and $E(u) = 1/\theta$. Here we have the marginal loglihood

$$m = \log L(\theta; y) = \log \theta - 2\log(\theta + y).$$

This gives the marginal ML estimator $\hat{\theta} = y$ and its variance estimator

$$\widehat{var}(\hat{\theta}) = -\{\partial^2 m/\partial\theta^2|_{\theta=\hat{\theta}}\}^{-1} = 2y^2.$$

Following the strategy of putting the constraints on random effects $E(w) = 1$ let us consider the model (17) first. Here because

$$\mu = E(y|w) = \theta/w,$$

the log link achieves additivity

$$\eta = \log\mu = \beta + v,$$

where $\beta = \log\theta$ and $v = -\log w$. This leads to the h-loglihood

$$h = l(\theta, v; y, v) \equiv \log f_\theta(y|v) + \log f(v) = -v - \log\theta - wy/\theta - w - v.$$

Suppose that θ and therefore β is known. The maximization $\partial h/\partial v = -2 + (y/\theta + 1)w = 0$ gives the BP

$$\hat{w} = 2\theta/(y + \theta) = E(w|y).$$

Here the BP is on the w scale. Then, the corresponding Hessian $-\partial^2 h/\partial w^2|_{w=\hat{w}} = 2/\hat{w}^2 = (y + \theta)^2/(2\theta^2)$ gives as an estimate for $var(\hat{w} - w)$

$$var(w|y) = 2\theta^2/(y + \theta)^2.$$

Now suppose that θ and therefore β is unknown. The joint maximization

$$\partial h/\partial v = -2 + (y/\theta + 1)w = 0 \text{ and } \partial h/\partial\theta = -1/\theta + yw/\theta^2 = 0$$

gives the ML estimator $\hat{\theta} = y$ and the empirical BP $\hat{w} = 2\hat{\theta}/(y + \hat{\theta}) = \widehat{E(w|y)} = \theta\widehat{E(u|y)} = 1$. Because there is only one random effect in the model the constraint $E(w) = 1$ makes $\hat{w} = 1 = E(w)$: for more discussion see Lee and Nelder (1996, 2004). Because

$$-\partial^2 h/\partial w^2|_{\theta=\hat{\theta},w=\hat{w}} = 2, \quad -\partial^2 h/\partial\theta^2|_{\theta=\hat{\theta},w=\hat{w}} = 1/y^2, \quad -\partial^2 h/\partial\theta\partial w|_{\theta=\hat{\theta},w=\hat{w}} = -1/y$$

we have an estimator

$$\widehat{var(\hat{\theta})} = 2y^2,$$

which is the same as that from the marginal loglihood. Now we have

$$\widehat{var(\hat{w} - w)} = 1 = var(w),$$

which reflects the variance increase caused by estimating θ; note that

$$\widehat{var(w|y)} = 2\theta^2/(y+\theta)^2|_{\theta=\hat{\theta}} = 1/2.$$

Here $-\partial^2 h/\partial v^2|_{w=\hat{w}} = 2$, so that h and $p_v(h)$ are proportional, showing that the joint maximization is a convenient tool to compute an exact ML estimator and its standard error estimates.

Suppose that we use the model (16) with an identity link. Now θ is a dispersion parameter appearing in $f_\theta(u)$ and the h-loglihood is given by

$$h = L(\theta, u; y, u) \equiv \log f(y|u) + \log f_\theta(u) = \log u + \log\theta - u(\theta + y).$$

Then, the equation $\partial h/\partial u = 1/u - (\theta + y) = 0$ gives $\tilde{u} = 1/(\theta + y)$. From this we get

$$p_u(h) \equiv \log\tilde{u} + \log\theta - \tilde{u}(\theta+y) - \frac{1}{2}\log\{1/(2\pi\tilde{u}^2)\} = \log\theta - 2\log(\theta+y) - 1 + \frac{1}{2}\log 2\pi,$$

which is proportional to the marginal loglihood m, and so yields the same inference for θ. Here $-\partial^2 h/\partial u^2|_{u=\tilde{u}} = 1/\tilde{u}^2 = (\theta+y)^2$ and thus h and $p_v(h)$ are no longer proportional, so that the joint maximization cannot give an exact ML estimator for dispersion parameters.

Schall's (1991) method is the same as Breslow and Clayton's (1993) PQL method for GLMMs. They are the same as the h-likelihood method, but ignore $\partial\tilde{u}/\partial\theta$ in the dispersion estimation (Lee and Nelder 2001*a*). Now suppose that the $\partial\tilde{u}/\partial\theta$ term is ignored in maximizing $p_u(h)$. Then we have the estimating equation

$$1 = \theta\tilde{u} = \theta/(\theta+y), \quad \text{for } y > 0$$

which gives an estimator $\hat{\theta} = \infty$. Thus, the term $\partial\tilde{u}/\partial\theta$ should not be ignored; if it is, it can result in a severe bias in estimation and a distortion of the standard error estimate; here, for example,

$$\widehat{var(\hat{\theta})} = \hat{\theta}^2 = \infty.$$

Similarly, in models (5) and (7) the joint maximization of $l(y, u|\lambda)$ does not provide a

valid inference for β as we saw, but $p_u(l(y,u|\lambda))$ does provide the ML estimator of β for both models. Thus, for model (8) even though we chose $l(y,w|\lambda)$ as the h-loglihood our inferential procedure provides equivalent inference to that from a model with a log-link. Thus, with proper use of h-likelihood it is possible to have meaningful inferences about both the random and fixed parameters.

For some non-linear random-effect models, such as those occurring in pharmacokinetics, the definition of h-likelihood is less clear, but we may still use $p_u(l(y,u))$ for inference about non-random parameters. Lee and Nelder (1996) noted that $p_u(l(y,u))$ is invariant, i.e. gives invariant inference, with respect to an arbitrary linear transformation of u. Even though the simplicity of the h-likelihood algorithm is lost, $p_u(l(y,u))$ provides a good inferential criterion because the Laplace approximation is often very accurate.

6.1 Discussion

There have been many alleged counterexamples similar to that of Bayarri *et al.* (1988), purporting to show that an extension of the Fisher likelihood to three objects is not possible. An important criticism is that we may get qualitatively different (i.e. non-invariant) inferences for trivial re-expressions of the underlying model. We see that defining the h-likelihood on the right scale avoids such difficulties. These complaints are caused by a misunderstanding of the h-likelihood framework and the wrong use of joint maximization to obtain all the parameter estimates. Another criticism has been about the statistical efficiency of the h-likelihood procedures (for example, Little and Rubin, 2002). An appropriate adjusted profile h-likelihood (APHL) should be used for estimation of dispersion parameters (Lee and Nelder, 1996, 2001). The h-likelihood is a natural way of making inferences about unobservables v. The definition of h in the proper scale of v and the use of APHLs give valid inferences. All the alleged counterexamples for missing data problems in Little and Rubin (2002; chapter 6.3) can be refuted similarly as in Section 6: for detailed discussion see Yun *et al.* (2005).

Hinkley (1979) and Butler (1986) introduced predictive likelihood for inferences of unobserved future observations. In missing data problemes and random-effect models, the APHL $p_v(h)$, eliminating random parameters v, is used for inferences about fixed θ. However, in prediction problems for an unobserved future observation v, various predictive likelihoods, eliminating fixed parameters θ, have been proposed for inferences about random v (Bjørnstad, 1990). Davison (1986) proposed to use the APHL $p_\theta(h)$, derived as an approximate predicted likelihood under a non-informative prior, which Butler (1990) called the modified profile predictive loglihood. Following Barndorff-Nielsen (1983) Bjørnstad (1990) suggested an approximation

$$\log f_\theta(z|y) = p_\theta(h) + \log\det(\partial\hat{\theta}/\partial\hat{\theta}_v),$$

where $\hat{\theta}$ is the ML estimator based upon y and $\hat{\theta}_v$ is the maximum h-logihood estimator.

If the number of predictands remains fixed as the number of y grows we have $\log\{\det(\partial\hat{\theta}/\partial\hat{\theta}_v)\} = O_p(1/n)$, so that we can also derive the predictive likelihood $p_\theta(h)$ as the approximate conditional likelihood, following Cox and Reid (1987). So, with h-likelihood perspectives, the predictive likelihood is an attempt to derive the APHL for predictions, eliminating all fixed parameters.

7 Advantages of extended likelihood framework

The difficulty in obtaining the two components $f_\theta(y)$ and $f_\theta(v|y)$ has limited the class of stochastic models that allow likelihood inference. Except for the limited conjugate family there are few models which allow explicit forms for these two components. The use of h-likelihood makes such limitations unnecessary, so that likelihood inference can be drawn from a much wider class of models. Furthermore, the extended likelihood framework preserves the advantages of the original framework.

7.1 Generality of application

HGLMs have become increasingly popular since the initial synthesis of GLMs, random-effect models, and structured-dispersion models was found to be extendable to include models for temporal and spatial correlations (Lee and Nelder 2001*a*, 2001*b*). Heterogeneity of means between clusters (the so-called between-cluster variation) can be modelled by introducing random effects in the mean. In HGLMs both fixed and random effects are allowed for the mean but only fixed effects for the dispersion. We have introduced double HGLMs (Lee and Nelder 2005), which allow both fixed and random effects not only for the mean but also for the dispersion. This means that heterogeneity of dispersion between clusters can be similarly modelled by introducing random effects in the dispersion. We now have a systematic way of generating heavy-tailed distributions for various types of data such as counts and proportions. This class will, among other things, enable models of types widely used in the analysis of financial data to be explored, and should give rise to new extended classes of models. The h-likelihood plays a key role in the synthesis of the inferential tools needed for these models.

7.2 Statistical efficiency of h-likelihood method

HGLMs have received increasing attention due to their wide applicability and ease of interpretation. However, the computation of the ML estimation of the parameters is a complex task. The marginal loglihood m, obtained by integrating out the random

effects, is in general analytically intractable. The computational problems are magnified when the random effects have a crossed design, where the data cannot be reduced to small independent clusters. For example, in the Salamander data marginal likelihood inference, based upon numerical integration using Gauss-Hermite quadrature is not feasible since a 120-dimensional integral is required. Thus, various approximate methods have been proposed by Schall (1991), Breslow and Clayton (1993), Drum and McCullagh (1993), Shun and McCullagh (1995), Lee and Nelder (1996, 2001), Lin and Breslow (1996) and Shun (1997). For binary data Noh and Lee (2004) showed numerically that the h-likelihood estimator has less bias than the other methods including MCMC-type methods: see also the simulation studies of Poisson and binomial models (Lee and Nelder 2001*a*), of frailty models (Ha *et al.*, 2001) and of mixed linear models with censoring (Ha *et al.*, 2002). We have not seen any method which outperforms the h-likelihood procedure, though we do not say that the current h-likelihood procedure is incapable of improvement.

7.3 Computational efficiency of h-likelihood method

The h-likelihood (13) gives a new definition of conjugate families (Lee and Nelder 2001*a*), showing that the likelihood for conjugate family for $\log f_\theta(v)$ takes the form of a GLM. It is sum of component likelihoods, $\log f_\theta(v)$ and $\log f_\theta(y|v)$, both representable as GLM likelihoods. This means that an extended class of models can be decomposed into component GLMs (Lee and Nelder 2001a, 2005) and these extended models can be fitted as an interconnected set of component GLMs. This greatly facilitates the development of model-checking techniques for the whole class (Lee and Nelder 2001*a*). A single algorithm, iterative weighted least squares, can be used throughout all these extended classes of models and requires neither prior distributions of parameters nor multi-dimensional quadrature. The h-likelihood plays a key role in the synthesis of the computational algorithms needed for this extended class of models.

This formulation means that a great variety of models can be fitted by a single algorithm and compared using extensions of standard GLM procedures. Thus 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. Furthermore various model assumptions can be checked by applying GLM model-checking procedures to the component GLM. This establishes, we believe, algorithmic *wiseness* in the sense of Efron (2003).

8 Conclusion

In general the computation of the ML and/or REML estimation of the parameters is a complex task due to the intractable integration to obtain the marginal loglihood. With the use of h-likelihood we can obtain ML and REML estimators by maximizing APHLs. However, still many believe that the marginal loglihood, without involving random effects, is the default loglihood. However, its use has always left a problem of inference about unobservable random variables (subject-specific inferences, Zeger *et al.*, 1988) and has restricted stochastic models to those having an explicit marginal likelihood. Thus, Bayesian methods have been extensively used for models without an explicit marginal likelihood, while likelihood inference is relatively less well developed, because the definition of likelihood for such inferences is not agreed.

We do not object to the use of marginal likelihood for inferences about fixed parameters, which in our approach appears as an APHL similar to the restricted likelihood. The restricted likelihood cannot allow inferences about fixed effects because they are eliminated. Similarly, the marginal likelihood cannot allow inferences about individuals, so that some other method must be used for this. As we use the marginal likelihood for inferences about β in the REML procedure it would be natural to use the h-likelihood for inferences about random effects. Thus, it is the h-likelihood that is fundamental, giving both marginal inference for fixed parameters and subject-specific inference for random or combined fixed and random parameters.

It is perhaps unfortunate that Bayesians, from Lindley and Smith (1972) onwards, seem to have made a take-over bid for all hierarchical models, implying that one has to be a Bayesian to deal with them. The availability of Markov-chain Monte Carlo, making models without an explicit marginal likelihood seem more easily handled via Bayesian computations, has appeared to justify this. By using h-likelihood, we may deal with models with random effects directly in a likelihood framework because there is an explicit analytic form of the likelihood. Furthermore inferences about random effects are possible without resorting to an empirical Bayesian framework. There seems to be no evidence that MCMC-type methods give better estimators than the h-likelihood method at least with binary data (Noh and Lee, 2004).

H-likelihood, as an extended likelihood, gives a powerful and practical framework for statistical inference; being a natural extension of Fisher likelihood to models with random parameters, it will become, we believe, widely used for inference for unobserved random variables. Nevertheless, it remains to be seen if any further generalizations can be made.

Acknowledgments We thank Sir David Cox, G. Casella, M. Crowder, M. Healy, J. Lawless, J. Lee, Y. Pawitan and S. Senn for their constructive comments, and P. McCullagh and C. McCulloch for providing stimulating examples and discussions. This

research was supported by a grant from Korea Research Foundation Grant (KRF-2003-070-C00008).

References

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

Barndorff-Nielsen, O. E. and Cox, D. R. (1996). Prediction and asymptotics. *Bernoulli*, 2, 319-340.

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

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, New York: Springer-Verlag.

Berger, J. O., and Wolpert, R. (1984). *The Likelihood Principle*. Hayward: Institute of Mathematical Statistics Monograph Series.

Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *J. Am. Statist. Ass.*, 57, 269-306.

Bjørnstad, J. F. (1990). Predictive likelihood principle: a review (with discussion). *Statist. Sci*., 5, 242-265.

Bjørnstad, J. F. (1996). On the generalization of the likelihood function and likelihood principle. *J. Am. Statist. Ass.*, 91, 791-806.

Box, M. J., Draper, N. R. and Hunter, W. G. (1970). Missing values in multi-response nonlinear data fitting. *Technometrics*, 12, 613-620.

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

Butler, R. W. (1986). Predictive likelihood inference with applications (with discussion). *J. R. Statist. Soc.* B, 48, 1-38.

Butler, R. W. (1990). Comment on "Predictive likelihood inference with applications" by Bjørnstad. *Statist. Sci*., 5, 255-259.

Carlin, B. P. and Louis, T. A. (2000). *Bayesian and Empirical Bayesian Methods for Data Analysis*. London: Chapman and Hall.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *J. R. Statist. Soc.* B, 32, 1-18.

Crouch, E. A. C. and Spiegelman, D. (1990). The evaluation of integrals of the form $\int_{-\infty}^{+\infty} f(t)\exp(-t^2)dt$: application to logistic-normal models. *J. Am. Statist. Ass.*, 85, 464-469.

Davison, A. C. (1986). Approximate predictive likelihood. *Biometrika*, 73, 323-332.

Dempster, A. P. N., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc.* B, 39, 1-38.

Drum, M. L. and McCullagh, P. (1993). REML estimation with exact covariance in the logistic mixed model. *Biometrics*, 49, 677-689.

Edwards, A. W. F. (1972). *Likelihood*. Cambridge: Cambridge University Press.

Efron, B. (2003). A conversation with good friends. *Statist. Sci.*, 18, 268-281.

Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, 1, 3-32.

Ha, I. D. and Lee, Y. (2005*a*). Multilevel mixed linear models for survival data. *Lifetime Data Analysis*, 11, 131-142.

— (2005*b*). Comparison of hierarchical likelihood versus orthodox best linear unbiased predictor approaches for frailty models, to appear in *Biometrika*, 42, 717-723.

Ha, I. D., Lee, Y. and Song, J. K. (2001). Hierarchical likelihood approach for frailty models. *Biometrika*, 88, 233-243.

— (2002). Hierarchical likelihood approach for mixed linear models with censored data. *Lifetime Data Analysis*, 8, 163-176.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection models. *Biometrics*, 31, 423-447.

Hinkley, D. V. (1979). Predictive likelihood. *Ann. Statist.*, 7, 718-728. Corr 8, 694.

Karim, M. R. and Zeger, S. L. (1992). Generalized linear models with random effects; salamander mating revisited. *Biometrics*, 48, 681-694.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion). *J. R. Statist. Soc.* B, 58, 619-678.

— (2001*a*). Hierarchical generalized linear models: A synthesis of generalised linear models, random-effect model and structured dispersion. *Biometrika*, 88, 987-1006.

— (2001*b*). Modelling and analysing correlated non-normal data, *Statistical Modelling*, 1, 3-16.

— (2004). Conditional and marginal models: another view (with discussion). *Statist. Sci.*, 19, 219-238.

— (2005). Double hierarchical generalized linear models (with discussion). to appear at *Appl. Statist.*

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.

Lindley, D. V. and Smith, A. F. M. (1972). Bayesian estimates for the linear model (with discussion). *J. R. Statist. Soc.* B, 34, 1-41.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data.* New York: John Wiley.

Mathiasen, P. E. (1979). Prediction functions. *Scand. J. Statist*., 6, 1-21.

Noh, M. and Lee, Y. (2004). REML estimation for binary data in GLMMs. Manuscript prepared for publication.

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

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference using Likelihood*. Oxford: Clarendon Press.

Pearson, K. (1920). The fundamental problems of practical ststistics. *Biometrika*, 13, 1-16.

Savage, L. J. (1976). On rereading R. A. Fisher (with discussion). *Ann. Statist.*, 4, 441-500.

Schall, R. (1991). Estimation in generalised linear models with random effects. *Biometrika*, 78, 719-727.

Searle, S. R., Casella, G. and McCulloch, C. E. (1992). *Variance Components*. New York: John Wiley and Sons.

Shun, Z. (1997). Another look at the salamander mating data: a modified Laplace approximation approach. *J. Am. Statist. Ass.*, 92, 341-349.

Shun, Z. and McCullagh, P. (1995). Laplace approximation of high-dimensional integrals. *J. R. Statist. Soc.* B, 57, 749-760.

Smyth, G. K. (2002). An efficient algorithm for REML in heteroscedastic regression. *J.Comp. Graph. Statist.*, 11, 1-12.

Vaida, F. and Meng, X. L. (2004). Mixed linears models and the EM algorithm in *Applied Bayesian and Causal Inference from an Incomplete Data Perspective*. Gelman, A. and Meng, X. L. (editors): John Wiley and Sons.

Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. *Emp. J. Exp. Agric.*, 1, 129-142.

Yun, S., Lee, Y. and Kenward, M. G. (2005). Using h-likelihood for missing observations. manuscript prepared for publication.

Zeger, S. L., Liang, K. Y. and Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 44, 1049-1060.