

Likelihood inference for models with unobservables: another view

Youngjo Lee and John A. Nelder

Abstract. There have been controversies among statisticians on (i) what to model and (ii) how to make inferences from models with unobservables. One such controversy concerns the difference between estimation methods for the marginal means not necessarily having a probabilistic basis and statistical models having unobservables with a probabilistic basis. Another concerns likelihood-based inference for statistical models with unobservables. This needs an extended-likelihood framework, and we show how one such extension, hierarchical likelihood, allows this to be done. Modelling of unobservables leads to rich classes of new probabilistic models from which likelihood-type inferences can be made naturally with hierarchical likelihood.

Key words and phrases: Hierarchical generalized linear model, unobservables, random effects, likelihood, extended likelihood, hierarchical likelihood.

1 Introduction

Fisher introduced the concept of likelihood in 1921 for inferences from statistical models involving two kinds of objects, namely observed random variables (data) and unknown fixed parameters. Pearson (1920) pointed out a limitation of Fisher likelihood for the prediction of unobserved future observations. Fisher's likelihood cannot be used to make inferences about unobservables. There has been an effort to extend likelihood inferences to models with unobservables by eliminating them via integration. However, with a few exceptions such as the copula (Joe, 1997), marginal distributions for

⁰*Youngjo Lee is a Professor, Department of Statistics, Seoul National University, Seoul, Korea. youngjo@snu.ac.kr. John A. Nelder is a Visiting Professor, Department of Mathematics, Imperial College, London, SW7 2AZ, UK. j.nelder@imperial.ac.uk.*

counts and proportions are not available in explicit forms, and this restricts the scope of the classical likelihood approach.

In longitudinal studies, generalized estimating equations (GEEs) are widely used. They give an estimation method for regression coefficients, constructed directly to describe marginal means, with the covariance structure regarded as contributing nuisance parameters only. However, GEEs cannot (generally) be integrated to obtain a likelihood function (McCullagh and Nelder, 1989) and therefore may not have a probabilistic or likelihood basis. These estimation methods for marginal (or population-average) means are often contrasted with conditional (or subject-specific) models, which include the modelling of unobservables. Jansen *et al.* (2006) reviewed the use of GEE methods and conditional models for analysis of missing data and discussed the choice between them. However, we believe that such a choice is inappropriate because the choice of an estimation method for a particular parameterization (marginal parameter) should not pre-empt the process of model selection. Recently, Lee and Nelder (2004) have shown that alleged differences in the behaviour of parameters between GEE methods and conditional models are based on a failure to compare like with like. We dislike the use of estimation methods without a probabilistic basis, because, for example, inferences for joint and conditional probabilities are not possible.

Recently, broad classes of new probabilistic models with unobserved random variables (unobservables) have been proposed, such as generalized linear models (GLMs) with random effects (Lee and Nelder, 1996), latent processes (Skrondal and Rabe-Hesketh, 2007), models for missing data (Little and Rubin 2002), prediction (Bjørnstad, 1990) and for potential outcomes in causality (Rubin, 2006) etc. In the statistical literature unobservables appear with various names such as random effects, latent processes, factor, missing data, unobserved future observations, potential outcomes etc. Random effects in the mean model have been proposed to account for within-subject correlation in longitudinal studies (Diggle *et al.*, 1996), for smooth spatial data (Besag and Higdon, 1999), for spline-type function fitting (Eilers and Marx, 1996), and for factor analysis (Bartholomew, 1987) etc., while random effects in the dispersion model (Lee

and Nelder, 2006a) can account for heteroskedasticity, giving heavy-tailed distributions that allow robust modelling (Noh and Lee, 2007a).

Modelling of unobservables is the key to these new models. However, because of difficulties in making likelihood inferences about unobservables, some authors use the Fisher likelihood for inferences about fixed unknown parameters, while for inferences about unobservables they use the empirical Bayesian (EB) approach, or full Bayesian (FB) inference. Recently, Zhao *et al.* (2006) have used a FB approach, which they claim to have an advantage over the frequentist version (EB) in that it is *computationally simpler* to obtain variance estimates of the random-effect estimates. (Note that the word ‘prediction’ has often been used to denote the estimation of random effects. However, we believe that it is clearer to use *prediction* when we estimate future observations (unobservables) and *estimation* for the estimation of random effects in the data already observed.) Discussing the controversy between Fisher and Neyman, Rubin (2005) maintained that models with unobservables arose most naturally in causal inference within a FB framework. From Lindley and Smith (1972) onwards, FB has become dominant for the analysis of models with unobservables. The availability of Markov-Chain Monte Carlo, which implements FB procedures, has made FB inferences popular.

By contrast we believe that modelling of unobservables is natural within an extended likelihood framework. Recently, for general inferences from models involving unobservables Lee and Nelder (1996) have proposed to use the hierarchical (or h-)likelihood. The h-likelihood plays a key role in the synthesis of the likelihood inferential tools needed for a broad class of new models having unobservables. The h-likelihood approach takes into account the uncertainty in the estimation of random effects, so that inferences about unobservables are possible without resorting to an EB framework.

In the next Section we review some models with unobservables and discuss related modelling issues. We review the h-likelihood procedure for the estimation of random effects and compare with the Bayesian approach in Section 3; likelihood inferences

from such models are demonstrated with examples in Section 4, followed by conclusions in Section 5.

2 How to model unobservables

Multivariate distributions for non-Gaussian models can be produced by probabilistic modelling of unobservables, without requiring explicit multivariate generalizations of non-Gaussian distributions. Using hierarchical likelihood, inferences from these new classes can be made.

2.1 HGLMs: random effects in the mean

HGLMs allow a synthesis of GLMs, random-effect models, and structured-dispersion models. Consider a GLM with random effects, where the response y follows the GLM, conditioning on random effects v :

$$\mu = \text{E}(y|v) \text{ and } \text{var}(y|v) = \phi V(\mu) \quad (1)$$

with a linear predictor

$$\eta = X\beta + Zv, \quad (2)$$

where $\eta = g(\mu)$ for some monotonic function $g(\cdot)$. When v are normal the models are called generalized linear mixed models (GLMMs). The use of other distributions for the random effects enriches the class of models. Lee and Nelder (1996) introduced HGLMs, in which the distribution of the random components is extended to an arbitrary conjugate distribution of a GLM family, with an appropriate link, not necessarily that of the conjugate pair. Above we suppress the indices to mean that our discussion covers various models having single or multiple random effects with nested, crossed, combined structures, etc. We write indices if necessary.

To allow various patterned associations among random effects Lee and Nelder (2001a) proposed to add an additional feature to HGLMs as follows: Let $v = Lr$

with r being random effects with a diagonal covariance matrix $\text{var}(r) = \Lambda$ to give

$$\text{var}(v) = \Sigma = L\Lambda L^t.$$

The last equation can be a spectral decomposition with an orthogonal matrix L or a Choleski decomposition with upper or lower triangular matrix L . Zhao *et al.* (2006) note that the full generality of the GLMM requires using general design matrices for both fixed and random components. With fixed L , not depending upon unknown parameters, we have models for longitudinal studies, intrinsic autoregressive models, various spline models etc. With parameter-dependent L we have random-slope models, autoregressive models, antedependence models, Markov-random-field models etc. (Lee and Nelder, 2001a). These models are also able to handle a great range of complications in regression-type analysis, for instance within-subject correlation in longitudinal data, scatterplot smoothing, generalized additive models, Kriging, function estimation and non-parametric regression models such as generalized additive models and varying-coefficient models (Zhao *et al.*, 2006).

Example 1: Consider the model from item-response theory (IRT) such that

$$\Pr(y_{ij} = 1|v_{ij}) = \frac{\exp(v_{ij} - \beta_j)}{1 + \exp(v_{ij} - \beta_j)},$$

where β_j is the intrinsic difficulty of the j th item and v_{ij} is the i th subject's ability for the j th item. If $v_{ij} = v_i$ with $v_i \sim N(0, \lambda)$ it becomes a one-parameter IRT model (Rasch, 1960). An appealing feature of this model is that items and subjects (examinees) can be placed on a common scale. Differences in both difficulty between items and ability of subjects is assumed to remain the same. In this model, for a given item, the probability of a correct response increases monotonically with ability as in Figure 1.

If $v_{ij} = r_i\alpha_j$ with $r_i \sim N(0, \lambda)$ and α_j fixed unknown, we have a two-parameter IRT model. Let $v_i = (v_{i1}, \dots, v_{ik})^t$ and $L_i = (\alpha_1, \dots, \alpha_k)^t$, giving

$$\text{var}(v_i) = \Sigma_i = L_i\Lambda L_i^t,$$

where $\Lambda = \lambda$ is a one-by-one matrix. This model allows for correlations among items for each subject. In this model α_j is called the discriminant parameter and $\beta_j^* = \beta_j/\alpha_j$ the difficulty parameter (Skrondal and Rabe-Hesketh, 2007). This two-parameter IRT model may lack the monotonicity property, in that one item can be easier than another for one subject, while being more difficult for another, this being described by the item-subject interaction $r_i\alpha_j$. This example shows how a particular modelling of the (singular) covariance matrix Σ_i can give an interesting interpretation of the parameters.

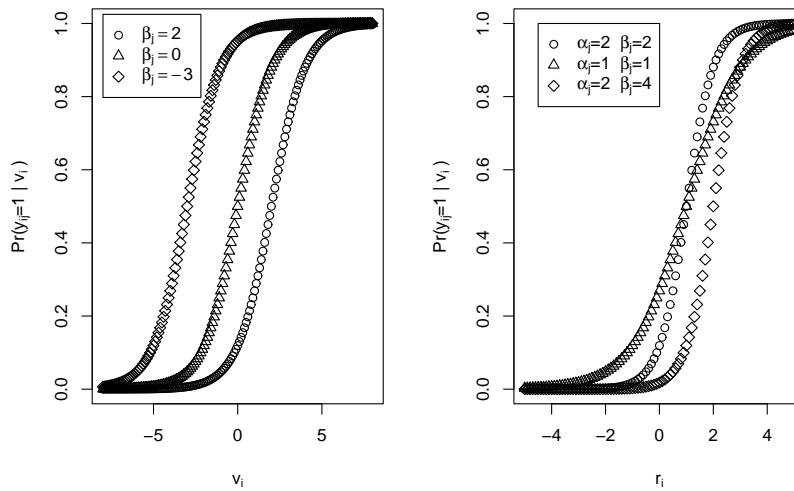


Figure 1: Curves of $\Pr(y_{ij} = 1|v_i)$ with respect to v (left) in a one-parameter IRT model and r (right) in a two-parameter IRT model

Example 2: When $v_t = \rho v_{t-1} + r_t$ with $\text{var}(r_t) = \lambda$ we have autoregressive random effects of order 1. When $\rho = 1$ we have the random-walk model, which gives a singular precision matrix. This random-walk model for temporal correlation has been extended to spatially correlated models via intrinsic autoregressive models with a singular fixed-precision matrix (Besag and Kooperberg, 1995). Splines can be viewed as smoothing via random effects which also have a singular fixed-precision matrix (Green and Silverman, 1994).

Example 3: Skrondal and Rabe-Hesketh (2004) proposed generalized linear la-

tent and mixed models (GLLAMMs) as a means of unifying factor models, linear structural-relations models, and covariate measurement-error models. They point out that GLLAMMs consist of two building blocks, a response model and a structural model. For the response model, they use the HGLM shown in equation (2). For the structural model, the random effect itself satisfies a regression model of the form

$$v = Bv + \Gamma w + r,$$

where B is a matrix of structural parameters relating the latent dependent variables to each other, Γ is a matrix of structural parameters relating the latent dependent variable to the latent explanatory variables, and r is a vector of disturbances. From this we have

$$v = (I - B)^{-1}\Gamma w + (I - B)^{-1}r.$$

Thus, the GLLAMMs can be represented as an HGLM with two random components

$$\eta = g(\mu) = X\beta + Zv = X\beta + ZL_1w + ZL_2r$$

where $L_1 = (I - B)^{-1}\Gamma$ and $L_2 = (I - B)^{-1}$. In GLLAMMs the parametrization using B , Γ , $\text{var}(w)$ and $\text{var}(r)$ gives a useful interpretation.

Other class of widely used models with unobservables are non-linear mixed-effect models in population pharmacokinetics and pharmacodynamics, models for missing data, and models for potential outcomes.

2.2 Random-effect models for the dispersion

Lee and Nelder (2006a) introduced double HGLMs (DHGLMs), which allow random effects for the dispersion. This gives a systematic way of generating heavy-tailed distributions for various types of data such as counts, proportions etc. Random effects in the mean affect the first two cumulants of the distribution of responses, while those in the dispersion affect the third and fourth cumulants, so that by allowing random effects in both mean and dispersion we can generate models with various patterns in the

first four cumulants. Castillo and Lee (2007) showed that DHGLMs provide a general treatment of Levy-process models in financial modelling, while Noh and Lee (2007a) showed that this new class allows robust modelling of GLM classes, with bounded influence. Yun and Lee (2006) showed how to model abrupt changes in the behaviour of schizophrenics. Glidden and Liang (2002) showed that sensitivity of estimators for β from HGLMs become more serious when the data form a selected sample. However, Noh *et al.* (2005) showed that by using a heavy-tailed distribution for the random effects, such a sensitivity in the estimators can be avoided.

2.3 Probabilistic and non-probabilistic methods

Without introducing random effects the GEE can be used to obtain maximum likelihood (ML) estimators when responses are normal. Estimates of regression coefficients from GEEs have been claimed to be consistent under various model misspecifications. It is often called the population-averaged model (Zeger *et al.*, 1998) or the marginal model (Jansen *et al.*, 2006) for a particular parameterization (regression coefficients for marginal means $E(y)$). For correlated non-normal responses, given a GEE $U(\beta_s) = \partial q / \partial \beta_s = 0$ (let us say), the mixed derivatives may not be the same (McCullagh and Nelder, 1989, p 337), *i.e.*

$$\partial^2 q / \partial \beta_s \partial \beta_r = \partial U(\beta_s) / \partial \beta_r \neq \partial U(\beta_r) / \partial \beta_s = \partial^2 q / \partial \beta_r \partial \beta_s;$$

if so there is no probabilistic model leading to the GEE $U(\beta_s) = 0$. Without such a basis the claim of consistency is meaningless: for more discussion see Crowder (1995) and Chaganty and Joe (2006).

It is of interest to study the class of marginal models, allowing estimating equations. Various marginal models have been proposed by Molenberghs and Lesaffre (1994), Molenberghs *et al.* (2007) and Heagerty and Zeger (2000). Heagerty and Zeger (2000) claimed that the parameter estimates from their marginal models were less sensitive to the misspecification of the distribution of random effects. Lee and Nelder (2004) showed that if one compared like with like the differences between the results from

the two models were not great. All that we can say is that certain parameterizations are less sensitive under certain probabilistic models, so that it could be recommended to use such a parameterization if it also met scientific requirements. For further controversies on parameterizations see Lindsey and Lambert (1998).

GEE is an estimating *method*, not a model. Thus, we do not believe that a useful comparison can be made between a probabilistic model such as a HGLM and an estimating method such as GEE. We see the analysis of data as consisting of three main activities: the first two are model fitting and model-checking which aim to find parsimonious well-fitting models, and together comprise model selection; the third is model prediction, where parameter estimates from selected models are used to predict quantities of interest and their uncertainties. In our view, inferences about margins and individual subjects' responses and a choice of an estimation method such as the GEE, ML etc, both belong to the prediction phase of the analysis.

In this paper we shall not consider GEE further because the method does not allow inferences about unobservables.

3 Extended likelihood versus Bayesian approaches

Besides the observed data and fixed unknown parameters in Fisher likelihood, an additional type of object, namely unobservable random variables v , is often of interest in making statistical inferences.

Example 4: Suppose that we have the number of epileptic seizures in an individual for five weeks, $y = (3, 2, 5, 0, 4)$. Suppose that these counts are i.i.d. from a Poisson distribution with mean θ . Now we want to have a predictive probability function for the seizure counts for the next week v . Here, $\hat{\theta} = (3 + 2 + 5 + 0 + 4)/5 = 2.8$, so that the plug-in technique gives the predictive distribution for the seizure count v of the next week:

$$f_{\hat{\theta}}(v = i|y) = f_{\hat{\theta}}(v = i) = \exp(-2.8)2.8^i/i!.$$

Pearson (1920) pointed out the limitation of Fisher likelihood using the plug-in method

because it cannot account for uncertainty in estimating θ .

Example 5: Suppose that the data Y are collected from the statistical model $f_\theta(Y; \theta)$. Suppose also that some of the intended observations in Y are unobservable because they are missing. We write $Y = (y_{obs}, y_{mis})$ for y_{obs} the observed and y_{mis} the missing components. Let r be missing data indicators such that

$$\begin{aligned} r_i &= 1, \text{ if } Y_i \text{ is missing,} \\ &= 0, \text{ if } Y_i \text{ is observed.} \end{aligned}$$

This leads to a probability function

$$f_\theta(Y, r; \theta) \equiv f_\theta(Y) f_\theta(r|Y).$$

Here $y = (y_{obs}, r)$ are the observed data and y_{mis} are the unobservables.

From these models, likelihood inferences can be made using the h-likelihood defined by

$$h = h(\theta, v) = \log f_\theta(y|v) + \log f_\theta(v) = \log f_\theta(y, v) = m + \log f_\theta(v|y), \quad (3)$$

where m is the marginal log-likelihood $m = \log f_\theta(y)$ with $f_\theta(y) = \int f_\theta(y|v) f_\theta(v) dv$. This is the (log) h-likelihood, which plays the same role as the log-likelihood m in Fisher's likelihood inference for models without unobservables. In forming the h-likelihood the choice of the scale for v is important (Lee *et al.*, 2006) because the mode and its curvature are used for inferences, as we shall discuss.

Throughout this paper we use $f_\theta(\cdot)$ to denote probability functions of random variables with fixed parameters θ ; the arguments within the brackets can be either conditional or unconditional. Thus, $f_\theta(y|v)$ and $f_\theta(v|y)$ have different functional forms though we use the same $f_\theta(\cdot)$ to mean probability functions with parameters θ .

3.1 Bayesian inferences

If we assume a prior $\pi(\theta)$ on parameters θ Bayesian inferences can be made. The posterior is

$$\pi(\theta, v|y) \propto \pi(y|v, \theta) \pi(v|\theta) \pi(\theta),$$

where $\pi(y|v, \theta) = f_\theta(y|v)$ and $\pi(v|\theta) = f_\theta(v)$. Here θ is also unobservable and is eliminated by integration. Let $\theta_{-i} = (\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_p)^T$. For Bayesian inferences various marginal or conditional posteriors have been used:

$$\pi(\theta|y) = \int \pi(\theta, v|y) dv,$$

$$\pi(\theta_i|y) = \int \pi(\theta, v|y) dv d\theta_{-i},$$

$$\pi(v_i|y) = \int \pi(\theta, v|y) dv_{-i} d\theta,$$

$\pi(v_i|y, \theta) = \int \pi(v|y, \theta) dv_{-i}$. In this paper full Bayesian (FB) inference is assumed to use the marginal posteriors $\pi(\theta_i|y)$ and $\pi(v_i|y)$, while empirical Bayesian (EB) inference (Morris, 1983) uses the conditional posteriors $\pi(v_i|y, \hat{\theta})$ where $\hat{\theta}$ are the ML estimators maximizing the likelihood $f_\theta(y) = \pi(\theta|y) = \int \pi(\theta, v|y) dv$ under the uniform prior $\pi(\theta) = 1$.

3.2 Adjusted profile h-likelihoods and likelihood inference

The likelihood principle of Birnbaum (1962) states that Fisher's marginal likelihood $f_\theta(y)$ carries all the (relevant experimental) information in the data about the fixed parameters θ , so that $f_\theta(y)$ should be used for inferences about θ : see also Berger and Wolpert (1984). For estimating fixed parameters θ we follow the likelihood principle by using the ML estimator from $f_\theta(y)$. We view the marginal likelihood as an adjusted profile likelihood eliminating nuisance unobservables v from the h-likelihood. However, the computation of ML estimators can be a complex task because of intractable integration. For example, in the Salamander data (McCullagh and Nelder, 1989) marginal-likelihood inference, based upon numerical integration using Gauss-Hermite quadrature, is not feasible since a 120-dimensional integral is required.

Let

$$\ell = \ell(\theta) = \log f_\theta(y) = \log \int \exp h dv$$

be the (log-) marginal likelihood. Let $l = l(\alpha, \psi)$ be a likelihood, either a marginal likelihood ℓ or an hierarchical likelihood h , with nuisance parameters α . Lee and Nelder

(2001) introduced a function $p_\alpha(l; \psi)$, defined by

$$p_\alpha(l; \psi) = [l - \frac{1}{2} \log \det\{D(l, \alpha)/(2\pi)\}]|_{\alpha=\tilde{\alpha}} \quad (4)$$

where $D(l, \alpha) = -\partial^2 l / \partial \alpha^2$ and $\tilde{\alpha}$ solves $\partial l / \partial \alpha = 0$. These $p(\cdot)$ functions define adjusted profile h-likelihoods (APHLs). If $\pi(\theta) = 1$ the Bayesian posterior is identical to the h-likelihood, $\pi(\theta, v|y) = f_\theta(y, v)$. Thus, APHLs can have a Bayesian interpretation; for example $p_{v_{-i}, \theta}(h; v_i)$ is the Laplace approximation to the marginal posterior $\pi(v_i|y)$, eliminating (v_{-i}, θ) by integration. When $\pi(\theta) = 1$ it is not a probability if the domain is the whole real line or the positive real line. However, as long as the marginal posterior is proper (finite) $\pi(v_i|y)$ would be considered as a valid posterior (Berger, 1985).

APHLs also allow a likelihood interpretation. Here $p_v(h; \theta)$ is the Laplace approximation to the marginal likelihood ℓ obtained by integrating over unobservables v (Lee and Nelder, 2001a); its maximum gives approximate (marginal) ML estimators for β . In likelihood inferences fixed parameters are eliminated by conditioning (if available) or profiling (in general). Now suppose that the parameters in a model can be divided into location parameters β and dispersion parameters σ^2 . Note that $p_\beta(\ell; \sigma^2)$ is an adjusted profile likelihood that approximates the conditional log-likelihood obtained by conditioning on the marginal ML estimator $\tilde{\beta}$ to eliminate the fixed unknown parameter β (Cox and Reid, 1987). A well-known exact example of this is the use of restricted likelihood in linear mixed models. Furthermore, $p_\theta(h; v)$ is Davison's (1986) predictive likelihood for v , eliminating nuisance fixed parameters θ . The APHL $p_{v_{-i}, \theta}(h; v_i)$ eliminates v_{-i} by integration and θ by conditioning on $\hat{\theta}$. When orthogonality does not hold between parameters we use a profile likelihood to eliminate nuisance parameters. To simplify the notation we sometimes suppress arguments, for example we use $p_v(h)$ instead of $p_v\{h(v, \beta, \sigma^2); \beta, \sigma^2\} = p_v(h; \beta, \sigma^2)$ if this does not lead to ambiguity.

Lee and Nelder (1996, 2001a, 2006a) proposed to maximize the h-likelihood h for the estimation of v , the marginal likelihood ℓ for the ML estimators for β , and the restricted likelihood $p_\beta(\ell)$ for the dispersion parameters σ^2 . Thus, our position is consis-

tent with the likelihood principle by using the marginal likelihood for inferences about θ . However, when ℓ is numerically hard to obtain, we propose to use adjusted profile h-likelihoods (APHLs) $p_v(h)$ and $p_{\beta,v}(h)$ as approximations to ℓ and $p_{\beta}(\ell)$; $p_{\beta,v}(h)$ approximates the restricted log-likelihood. Second-order Laplace approximations may sometimes be useful to improve accuracy.

Many numerical studies on h-likelihood have shown that this development gives practically satisfactory estimates of parameters in many models where the ML estimators are hard to compute. For binary data Noh and Lee (2007b) showed numerically that the h-likelihood estimator for θ has less bias and mean square error than various other methods developed by Schall (1991), Breslow and Clayton (1993), Drum and McCullagh (1993), Shun and McCullagh (1995), Lin and Breslow (1996) and Shun (1997): see also the simulation studies of frailty models (Ha and Lee, 2005) and of mixed linear models with censoring (Ha *et al.*, 2002). In the salamander data, among other methods considered, the MCEM of Vaida and Meng (2004) gives the closest estimates to the h-likelihood estimators.

Little and Rubin (2002) provided an extensive review of the analysis of missing data and claimed that h-likelihood methods were inappropriate for the estimation of θ in missing-value settings such as that in Example 5. They appear wrongly to have equated h-likelihood estimation to a joint maximization of mean and dispersion parameters. Yun *et al.* (2007) showed, in contrast to this assertion, that when applied appropriately h-likelihood methods are both valid and efficient in such settings. In non-linear mixed-effect models the h-likelihood can also improve on existing methods (Noh and Lee, 2008).

3.3 APHLs versus marginal posteriors

In the Bayesian approach, simulation techniques such as MCMC are often used to compute the marginal posteriors. Consider the Epil example of the OpenBUGS manual, volume 1 (Thomas *et al.* 2006). The data come from a clinical trial of 59 epileptic

patients. Each patient i is randomized to a new drug ($T_i = 1$) or a placebo ($T_i = 0$). The observations for each patient y_{i1}, \dots, y_{i4} are the number of seizures during the 2 weeks before each of four visits. The covariates are age (A_i), the baseline seizure counts (B_i), and an indicator variable for the fourth clinic visit ($V4$). Consider the HGLM

$$\eta_{ij} = \beta_0 + \beta_B \log(B_i/4) + \beta_T T_i + \beta_{T \times B} T_i \times \log(B_i/4) + \beta_A A_i + \beta_V V4 + v_i + w_{ij},$$

using centered covariates with $v_i \sim N(0, \sigma_v^2)$ and $w_{ij} \sim N(0, \sigma_w^2)$. In discussing the paper by Rue *et al.* (2008) on Bayesian inferences based on priors $\sigma_v^{-2}, \sigma_w^{-2} \sim \text{gamma}(0.001, 0.001)$, Lee showed Figure 2 (of this paper) for the marginal posteriors, $\pi(v_1|y)$, $\pi(\beta_T|y)$ and $\pi(\sigma_v^2|y)$, from OpenBUGS (Thomas *et al.*, 2006) and the corresponding APHLs, $p_{v-1, w, \theta}(h; v_1)$, $p_{v, w}(h; \beta_T, \hat{\theta}(\beta_T))$ and $p_{v, w, \beta}(h; \sigma_v^2, \hat{\sigma}_w^2(\sigma_v^2))$, where $\hat{\theta}(\alpha)$ are the ML estimators of remaining β and the REML estimators for the dispersion parameters at $\beta_T = \alpha$ and $\hat{\sigma}_w^2(\alpha)$ is the REML estimators of σ_w^2 at $\sigma_v^2 = \alpha$. Figure 2 shows almost identical plots for both random and fixed effects. However, the plots for the dispersion components are different because Rue *et al.*'s (2008) inverse-gamma prior is informative. This leads to biases when dispersion parameters are not random but are fixed unknowns, as in disease mappings (Jang *et al.*, 2007). Thus, without MCMC samplings similar information could be obtained from the extended likelihood unless the assumed prior is informative. Thus, likelihood inferences can be made without the necessity of inventing priors for parameters.

4 Likelihood inference for unobservables

The extended likelihood principle of Bjørnstad (1996) shows that extended likelihood, of which h-likelihood is a special case, carries all the information in the data about the unobserved quantities v and θ . Bedrick and Hill (1999) studied the use of extended likelihood as a summary function for unobservables. In this paper we discuss its use as an estimating tool.

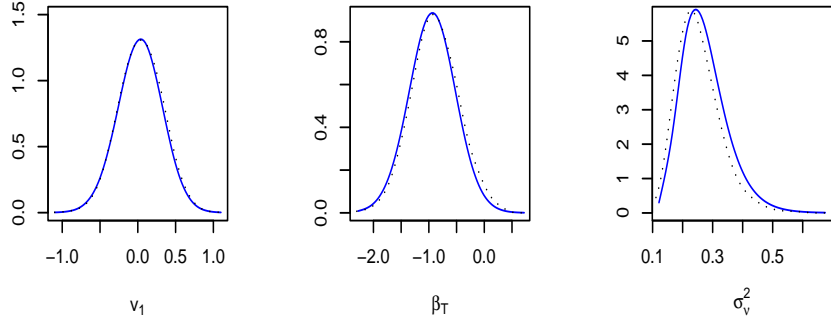


Figure 2: The marginal posteriors (\cdots) versus APHLs ($-$)

Consider the prediction problem in Example 4, where the plug-in technique $f_{\hat{\theta}}(v = i) = f_{\hat{\theta}}(v = i|y) = \pi(v = i|y, \hat{\theta})$ can be viewed as the EB. With Jeffreys' prior, $\pi(\theta) \propto \theta^{-1/2}$, the resulting marginal posterior $\pi(v|y)$ gives a predictive probability with higher probabilities for larger y . Pawitan (2001) considered the h-likelihood, proportional to

$$f_{\theta}(3, 2, 5, 0, 4, v) = \exp(-6\theta)\theta^{3+2+5+0+4+v}/(3!2!5!0!4!v!).$$

Here $\hat{\theta}(v) = ((3 + 2 + 5 + 0 + 4 + v)/6)$. Then, the normalized profile likelihood $f_{\hat{\theta}(v)}(3, 2, 5, 0, 4, v)$ gives the predictive distribution of Mathiasen (1979), almost identical to Pearson's but without assuming a prior on θ (Figure 3): for more discussion see Bjørnstad (1990). This example shows that standard methods for likelihood inferences can be used for the prediction problem. In the next Section we illustrate how to use standard likelihood methods to overcome a drawback of EB method.

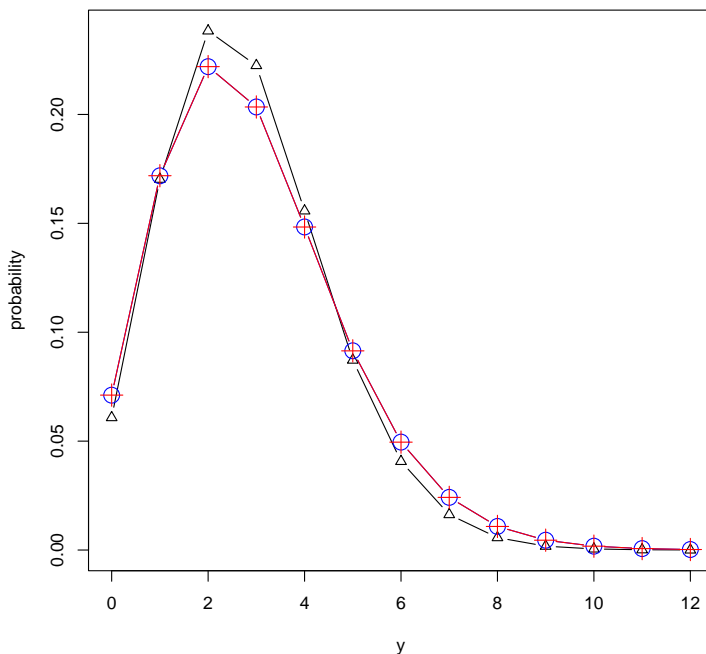


Figure 3: Predictive density of the number of seizure counts: Plug-in method (\triangle), Bayesian method (\circ) and h-likelihood method (+).

4.1 EB versus h-likelihood methods

Because the Fisher likelihood $f_{\theta}(y)$ in (3) does not involve v , the other component (the conditional posterior) $f_{\theta}(v|y) = \pi(v|y, \theta)$ seems to carry all the information in the data about the unobservables. Thus, an inference would be based solely upon the estimated posterior

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

where $\hat{\theta}$ are usually the ML estimators (Carlin and Louis, 2000). Using $f_{\hat{\theta}}(v|y)$ to make inferences about v is naive and Bjørnstad (1990) has shown how badly it performs in measuring the true uncertainty in estimating v . Note that maximization of the h-likelihood (3) yields EB-mode estimators for v , without computing $f_{\theta}(v|y)$. However,

the Hessian matrix based upon the estimated posterior $f_{\hat{\theta}}(v|y)$ gives a naive variance estimate for the prediction \hat{v} because it does not properly account for the uncertainty caused by estimating θ . Note that the marginal posterior variance is

$$\text{var}(v_i|y) = E_{\theta|y}[\text{var}(v_i|y, \theta)] + \text{var}_{\theta|y}[E(v_i|y, \theta)]. \quad (5)$$

Carlin and Gelfand (1991) noted that the naive EB variance estimate only approximates the first term in the equation above. Laird and Louis (1987) and Carlin and Gelfand (1990) proposed to use bootstrap method to estimate the second term. In this paper the FB method uses the marginal posterior $\pi(v_i|y)$.

Up to now most studies on h-likelihood methods have been about the efficiencies of parameter estimates. Here we discuss how to compute the variance of estimated random effects. We see that inferences about random effects cannot be made by using only $f_{\theta}(v|y)$, 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 in estimating θ ; it is the other component $f_{\theta}(y)$ which carries the information about this. By using the h-likelihood, complete likelihood inferences can be made not only for θ but also for v and their combinations.

Given θ let $\hat{v}(\theta)$ be a random-effect estimator solving $\partial h / \partial v = 0$. As a variance of random-effect estimators Booth and Hobert (1998) recommend using the conditional mean square error (CMSE) defined by

$$CMSE(v) = E\{(\hat{v}(\hat{\theta}) - v)(\hat{v}(\hat{\theta}) - v)'|y\} = \text{var}_{\theta}(v|y) + D(\theta), \quad (6)$$

where $\text{var}_{\theta}(v|y) = E\{(\hat{v}(\theta) - v)(\hat{v}(\theta) - v)'|y\}$ and $D(\theta) = E\{(\hat{v}(\hat{\theta}) - \hat{v}(\theta))(\hat{v}(\hat{\theta}) - \hat{v}(\theta))'|y\}$ is the inflation of the CMSE caused by estimating θ . The EB estimator, the inverse of the Hessian matrix from $\log f_{\theta}(v|y)$, gives an estimator for the first term $\text{var}_{\theta}(v|y)$ in (6). Thus, it could give severe underestimation if $D(\theta)$ is large. Lee and Nelder (1996) noted that in HGLMs (2) the location parameters (v, β) and dispersion parameters $\sigma^2 = (\phi, \Sigma)$ are orthogonal, so that we need consider only the variance

inflation caused by estimating β . The Hessian matrix of β and v is given by

$$I(\beta, v) = - \begin{pmatrix} \partial^2 h / \partial \beta \partial \beta' & \partial^2 h / \partial \beta \partial v' \\ \partial^2 h / \partial v \partial \beta' & \partial^2 h / \partial v \partial v' \end{pmatrix}. \quad (7)$$

Here the EB variance estimator is given by $-(\partial^2 h / \partial v \partial v')^{-1}|_{\theta=\hat{\theta}}$. Lee and Ha (2008) showed that in general the inverse of the Hessian matrix (7) gives an approximation to the CMSE (6). Before we discuss the general use of this method we investigate a simple example which shows issues related to this problem.

4.2 Bayarri's example

Bayarri *et al.* (1988) tried to show by an example that likelihood inference is not possible for general models with unobservables. Suppose that there is a single fixed parameter θ , a single unobservable random quantity u and a single observable quantity y . An unobserved random variable u has a probability function

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

and an observable random variable y has conditional probability function

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

free of θ . Besides $f(y|u)$, they considered two additional possibilities for an extended likelihood for models with these three kinds of objects:

$$\begin{aligned} f_{\theta}(y) &= \frac{\theta}{(\theta + y)^2}, \\ f_{\theta}(y, u) &= u\theta \exp\{-u(\theta + y)\}. \end{aligned}$$

The marginal log-likelihood $m = \log f_{\theta}(y)$ gives the ML estimator for θ but is totally uninformative about the unknown value of u . The conditional likelihood $f(y|u)$ is uninformative about θ and loses the relationship between u and θ reflected in $f_{\theta}(u)$. Finally, the extended likelihood $f_{\theta}(y, u)$ yields, if maximized jointly 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 complete inferences, so that Bayes is the only method for inferences from general models.

The h-(log)-likelihood is given by

$$h = \log f_{\theta}(y, v) = \log f_{\theta}(y, u) + \log |du/dv| \equiv 2v + \log \theta - u(\theta + y),$$

where $v = \log u$ with v being the canonical scale in which the joint maximization of h with respect to θ and u gives the ML estimator of θ (Lee *et al.*, 2006). Suppose that the marginal likelihood is hard to obtain. The Laplace approximation is proportional to $m = \log f_{\theta}(y)$ and gives the ML estimator $\hat{\theta} = y$ and its variance estimator

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

Given θ , the estimating equation $\partial h / \partial u = 0$ gives the best estimator of u (Robinson, 1991)

$$\hat{u}(\theta) = \text{E}(u|y) = \frac{2}{\theta + y}$$

from which we have

$$\hat{u}(\hat{\theta}) = \frac{2}{\hat{\theta} + y} = \frac{1}{y}.$$

Furthermore, we have

$$I(\theta, \hat{u}(\theta)) = - \begin{pmatrix} \partial^2 h / \partial \theta^2 & \partial^2 h / \partial \theta \partial u \\ \partial^2 h / \partial u \partial \theta & \partial^2 h / \partial u^2 \end{pmatrix} = \begin{pmatrix} 1/\theta^2 & 1 \\ 1 & (y + \theta)^2/2 \end{pmatrix}.$$

Note here that

$$\text{var}_{\theta}(u|y) = \text{E}\{(\hat{u}(\theta) - u)^2|y\} = 2/(y + \theta)^2,$$

so that EB gives $\widehat{\text{var}}_{\theta}(u|y) = 1/(2y^2)$. Here $D(\theta) = \text{E}\{[1/y - 2/(\theta + y)]^2|y\} = (y - \theta)^2 / \{y(y + \theta)\}^2 = (\hat{\theta} - \theta)^2 / \{y(y + \theta)\}^2$, so that, following Booth and Hobert (1998), if we estimate $(\hat{\theta} - \theta)^2$ by $\text{var}(\hat{\theta})$ we have $\widehat{D}(\theta) = 2y^2/4y^4 = 1/(2y^2)$. Thus, the estimator for the CMSE is $1/y^2$, which can be obtained from the corresponding element in the Hessian matrix $I(\hat{\theta}, \hat{u}(\hat{\theta}))$. An alternative justification is that the h-likelihood variance estimator is estimating the unconditional mean-square error because $\text{E}\{(\widehat{\hat{u}(\hat{\theta})} - u)^2\} = 1/y^2$ from $\text{E}\{(\hat{u}(\hat{\theta}) - u)^2\} = 1/\theta^2$ (Lee *et al.* 2006, page 116).

With this small example we illustrate how the h-likelihood gives complete likelihood inferences, giving the ML inference for θ and improved EB inference by accounting for the uncertainty caused by estimating θ .

4.3 H-likelihood inferences about v

The example shows that between extended likelihoods $f_\theta(y, u)$ and $f_\theta(y, v)$ the mode of the h-likelihood $f_\theta(y, v)$ gives a meaningful estimator for v , while that of $f_\theta(y, u)$ gives a meaningless one. Given that extended likelihoods should serve as the basis for statistical inferences of a general nature, we want to find a particular scale whose mode gives meaningful inferences about unobservables. Under the canonical scale the example shows that the mode gives the best estimator of u $E(u|y)$. However, the canonical scale does not exist in general. In HGLMs Lee and Nelder (2005) showed that maintaining invariance of inference from extended likelihood for trivial re-expressions of the underlying model leads to a unique definition of the h-likelihood; we call this the weak canonical scale in which v appears in the linear predictor.

In Section 3.3 we showed that APHLs are often similar to marginal posteriors. Given (marginal) posteriors, a Bayesian would use a decision-theoretic approach to choose estimators, while we use the mode of the h-likelihood (an extended likelihood on a particular scale) or its APHLs. Thus, the choice of the scale in defining the h-likelihood is important to guarantee the meaningfulness of the mode estimation. Lee and Ha (2008) showed that the standard error estimators from the Hessian matrix (7) give the first-order approximation to (5) with $\pi(\theta) = 1$ (Kass and Steffey, 1989) and to the CMSE (Booth and Hobert, 1998). Let $w = k(u)$ for some monotone function $k(\cdot)$. Ha and Lee (2006) showed conditions when the approximation becomes better. One such condition is that $w|y$ follows the normal distribution. In GLMMs when v is normal we may expect $v|y$ to be approximately normal. If normal the Laplace approximation is exact; we expect that proposed h-likelihood method works well. Figure 2 shows how to check the normality of the conditional distribution by using the APHL.

4.3.1 Analysis of the BC Infant Mortality Data

For disease mapping Leroux *et al.* (1999) and MacNab *et al.* (2004) considered the conditional autoregressive (CAR) model for the relative risk v_i , which satisfies $v \sim N(0, \Sigma)$, where $\Sigma = \sigma^2 D^{-1}$, $D = \lambda Q + (1 - \lambda)I$, σ^2 is a dispersion parameter reflecting the overall heterogeneity of the underlying risks and λ is a dispersion parameter for the spatial autocorrelation, $\lambda \in [0, 1]$. The neighborhood matrix Q has the j th diagonal element equal to the number of neighbors of the corresponding local region, while the off-diagonal elements in each row are equal to -1 if the corresponding regions are neighbors and 0 otherwise.

The data consist of the number of infant deaths and aggregated mid-year estimates of the population sizes of infants for 79 local health areas. Population size n_i varies from 123 to 52856. For these data Lee *et al.* (2007) compared inferences from the h-likelihood with the full Bayes (FB) analysis. For the FB approach, they set priors $\beta_i \sim N(0, 1/0.00001)$ and $\sigma^{-2} \sim \text{gamma}(0.0001, 0.0001)$. Initial values are set as $\sigma^2 = 1$, $\beta_i = 0$, and $v_i = 0$ and they obtain a posterior sample of 10,000, setting thinning at 10 using WinBUGS (MacNab *et al.*, 2004). The coverage probability is calculated by 95% Wald confidence intervals, based upon asymptotic normality, for the relative risks (v) using EB and h-likelihood, and in the FB method by equal-tail 95% credible intervals, the interval between the 2.5th and 97.5th percentiles of the posterior distribution as given by WinBUGS. For the FB method we use 10,000 iterations after a burn-in of 2000.

Lee *et al.* (2007) did a simulation study, assuming n_i and neighborhood structures identical to those in the BC infant mortality; the data were generated based on (1.1) and (3.1) with $\beta = -4.920$, $\sigma^2 = 2$ and $\lambda = 0.62$. Using a graph similar to Figure 4, they showed that the EB coverage probability decreases dramatically as the population size n_i increases, but that both the h-likelihood and FB methods improve the EB method substantially by accounting for the uncertainty in estimating fixed parameters. However, the coverage probability of FB also decreases as n_i increases, while

the h-likelihood maintains the stated level of confidence. When n_i becomes larger the priors for the dispersion parameters in the FB may cause problems in frequentist coverage probability. The h-likelihood procedure maintains the frequentist coverage probabilities better in this problem. The h-likelihood method is superior to Ainsworth and Dean's (2006) penalized quasi-likelihood (Lee *et al.*, 2007) for spatial GLMMs and Ma and Jorgensen's (2007) orthodox BLUP method (Lee and Ha, 2008) for non-normal Tweedie models.

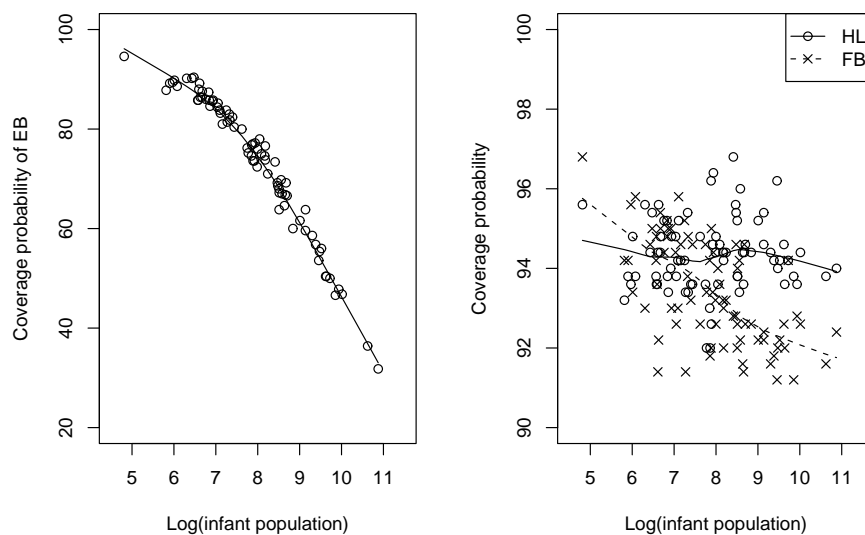


Figure 4: Coverage probabilities of the EB (left) , FB and h-likelihood (right) methods with respect to population size in the infant mortality data.

4.4 Inferences and model identifiability

The joint model for $f_{\theta}(y, v)$ leads to a marginal model $f_{\theta}(y)$ for the observed data. We regard $f_{\theta}(y, v)$ as the fundamental model, from which the marginal model can be made. However, different models for unobservables in $f_{\theta}(y, v)$ can lead to the same marginal model $f_{\theta}(y)$, so that care is necessary in making inferences about unobservables. Some model assumptions can be checked from the data, while some cannot. This could

be an advantage of objective inference with the likelihood, where uncheckable model assumptions cannot be identifiable. In Bayesian analysis priors can give information on unidentifiable model assumptions, so that it is hard to know whether the information is coming entirely from the uncheckable priors.

In the modelling of incomplete data we may assume the missing data to be missing not at random (MNAR) or assume random missingness (MAR). Here assumptions for the missing mechanism cannot be checked by using observed data (Rubin 2006). Molenberghs *et al.* (2007) further showed that an empirical distinction between MAR and MNAR is not possible because each MNAR model fit to a set of observed data can be reproduced exactly by its counterpart. Such a pair of models will produce identical estimates for the observed data, but give different estimates for the unobservables (missing data). Assumptions about unobservables (missing data) are not checkable without additional information. Unless we have a side-study to determine whether the observation process depends on what would be observed, all we have is a model-based assessment. As a referee has pointed out, it will contain some unverifiable assumptions.

In HGLMs model assumptions for unobservables are often verifiable, i.e. checkable by using the data because the unobservables are latent variables for observed data. Consider the one-way random-effect model

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

where $v_i \sim N(0, \lambda)$ and $e_{ij} \sim N(0, \phi)$, with v_i and e_{ij} uncorrelated. With more than one observation in each group the within-group error components v_i and e_{ij} are separately estimable, providing variance-component estimates for the dispersion parameters. Here model parameters ϕ and λ connect the observed data and unobservables. Lee and Nelder (2006b) showed that if there are different random-effect models giving the same induced marginal model for the observed data, then the h-likelihood inferences give equivalent inferences for equivalent pairs of objects, including unobservables. This model leads to a marginal model, namely the compound-symmetric

model:

$$Y_i \sim N(\mathbf{1}\beta, \lambda J_{n_i} + \phi I_{n_i}).$$

A compound-symmetry model with negative correlation $\lambda < 0$ is perfectly natural in a variety of settings (Nelder, 1954), which can be tested by the marginal likelihood (or APHL). Such a model can be covered by HGLMs if we allow a negative variance, but then many unanswered questions arise, such as estimability of random effects etc.; these require further research.

Wilk and Kempthorne (1957) and Cox (1958) studied the randomization theory of the Latin square, paying particular attention to the effects on the interpretation of the conventional analysis of variance (ANOVA) of the absence of unit-treatment additivity, a point first raised by Neyman (1935). Consider a model for the Latin-square design

$$y_{ij(k)} = \mu + r_i + c_j + \tau_k + (rc)_{ij} + (rt)_{ik} + (ct)_{jk} + e_{ij(k)}. \quad (8)$$

Suppose that the main effects are regarded as fixed. When the interactions $(rc)_{ij}$, $(rt)_{ik}$, $(ct)_{jk}$ are fixed a test for the main effect is irrelevant, because it makes no sense to postulate that either of the two main effects is null when their interaction is not assumed zero (Nelder, 1994). However, if the interactions are regarded as random the associated main effects can be tested without any difficulty from the ANOVA table. Permutation from a finite population is a way of generating distributions for random effects. Wilk and Kempthorne (1957) put constraints $\sum_i (rc)_{ij} = \sum_j (rc)_{ij} = \dots = \sum_k (ct)_{jk} = 0$. Nelder (1994) pointed out that such constraints make no sense either with fixed or random effects. With fixed effects the choice of constraints to give the least-square equations a solution is essentially arbitrary. However, with random effects symmetric constraints on estimates of the parameters of the form $\sum_i \widehat{(rc)}_{ij} = \sum_i \widehat{(rc)}_{ij} = \dots = \sum_k \widehat{(ct)}_{jk} = 0$ arise naturally (Lee and Nelder, 1996, 2005). However, here only fractions of combinations are used to make the combined error component $v_{ij(k)} = (rc)_{ij} + (rt)_{ik} + (ct)_{jk} + e_{ij(k)}$ to form a sum of independent errors. Thus, model (8) gives an identical marginal model to the conventional model for Latin squares with

main effects only

$$y_{ij(k)} = \mu + r_i + c_j + \tau_k + e_{ij(k)}^*. \quad (9)$$

From Lee and Nelder (2006b) the two models lead to identical inferences about both fixed parameters and random effects, giving $\hat{e}_{ij(k)}^* = \hat{v}_{ij(k)}$. Thus, in (8) individual error components cannot be separated by the observed data. If a method can identify individual components then it must be based upon uncheckable model assumptions such as priors. Consider the following model

$$y_{ij(k)} = \mu + r_i + c_j + \tau_{ij(k)} + e_{ij(k)}, \quad (10)$$

where $\tau_{ij(k)} = \tau_k + (rt)_{ik} + (ct)_{jk}$ and $(rt)_{ik}$ and $(ct)_{jk}$ are random with zero means. This model assumes unit-treatment interaction and can be interpreted to have the average treatment effects such that

$$E(\tau_{ij(k)}) = \tau_k.$$

Then we can test that the average treatment effects are the same (Lee and Nelder, 2002). Thus, with unobservables there are different methods of interpretation: we may consider $(rt)_{ik}$ and $(ct)_{jk}$ to be either error components or random treatment-unit interactions. These give equivalent inferences for equivalent quantities.

4.5 Discussion

There have been many alleged examples similar to that of Bayarri *et al.* (1988) and Little and Rubin (2002, Chapter 6.3), purporting to show that an extension of the Fisher likelihood to three kinds of objects is not possible. Lee and Nelder (2005) refute those of Bayarri *et al.* and Yun *et al.* (2007) those of Little and Rubin. These complaints are, we believe, resolved by the h-likelihood framework. Zhao *et al.* (2006) claimed that the Bayesian analysis is computationally simpler for obtaining variance estimators for the random-effect estimates compared with its frequentist counterpart; however with the extended likelihood framework this may not be so, at least in the analysis of the disease-mapping areas in Section 4.3.1.

The h-likelihood (3) gives a new definition of conjugate families (Lee and Nelder 2001a), showing that the likelihood for a conjugate family for $\log f_{\theta}(v)$ takes the form of a GLM. It is the 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, 2006a) and that 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 2001a). A single algorithm, iterative weighted least squares, can be used throughout all this extended class 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 appropriate component GLMs. This establishes, we believe, algorithmic *wiseness* in the sense of Efron (2003).

5 Conclusion

We have shown that a broad class of new models with wide applications can be generated by the probabilistic modelling of unobservables. There has been an attempt using the GEE method to make inferences from general non-normal multivariate models without modelling unobservables. It pre-empts model selection by claiming to make inferences about population averages or marginal means. We do not disagree with the need to make marginal predictions after choosing a model, but believe that such a need does not require, and indeed should not use, prediction methods at the model-selection

stage. We dislike the pre-emption of the model selection stage by a particular prediction method. Furthermore, these population, marginal and subject-specific averages are parameterizations in the probabilistic model. When a prediction method lacks a probabilistic model basis it is not possible to connect these parameters and compare them.

We do not object to the use of Fisher's likelihood for inferences about fixed parameters. The Fisher likelihood framework has advantages such as generality of application, statistical and computational efficiency etc. and we agree with its use. However, it cannot deal with inferences from models having unobservables because there is always a problem of inference about those unobservables. H-likelihood gives a powerful and practical framework for statistical inference of general model class with unobservables, maintaining the advantages of the original likelihood framework for fixed parameters. We believe that more new classes of models will be developed and that the h-likelihood will become widely used for inference from them.

The h-likelihood uses the mode and its curvature for inferences about unobservables. Thus, in defining the h-likelihood the scale of unobservables must be carefully chosen to make a valid inferences. The (weak) canonical scale in HGLMs leads to an invariance of a certain extended likelihood. However, in general the validity of such a scale has not been established. The conditional normality in Section 4.3 would be a promising condition to determine the scale, which can be checked by plotting the APHL. Further studies are required on the scale in defining the h-likelihood under general situations beyond DHGLMs. For fixed parameter estimation we use the marginal likelihood. But it is often hard to compute, so that we have proposed using the Laplace approximation. However, this approximation gives non-negligible biases in binary data. We have found that the second-order approximation is effective in eliminating such biases. However, it becomes very hard to implement as the number of random components increases. So it would be of interest to find an approximation which can be implemented under general situations.

ACKNOWLEDGMENTS

The authors thank Professors Jan Bjørnstad, Martin Crowder, Harry Joe, Jaeyong Lee, Yudi Pawitan and Roger Payne for their helpful comments. This work was supported by Brain Korea 21.

REFERENCES

- Ainsworth, L. M. and Dean, C. B. (2006) Approximate inference for disease mapping. *Comp. Statist. Data Anal.* **50**, 2552-2570.
- Bartholomew, D. J. (1987) *Latent variable models and factor analysis*. Oxford University Press, Oxford.
- 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, Springer, New York.
- Bedrick, E.J. and Hill, J.R. (1999). Properties and Applications of the Generalized Likelihood as a Summary Function for Prediction Problems. *Scandinavian Journal of Statistics*, **26**, 593-609.
- Berger, J. O. (1985), *Statistical decision theory and Bayesian analysis*. Springer, New York.
- Berger, J. O. and Wolpert, R. (1984), *The Likelihood Principle*. Institute of Mathematical Statistics Monograph Series, Hayward.
- Besag, J. and Higdon, P. (1999) Bayesian analysis of agricultural field experiments (with discussion). *J. R. Statist. Soc. B.*, **61**, 3-66.
- Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, **82**, 783-746.
- 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. Amer. Statist. Ass.*, **91**, 791-806.
- Booth J. G. and Hobert, J. P. (1998). Standard errors of prediction in generalized linear mixed models. *J. Amer. Statist. Assoc.*, **93**, 262-272.
- Breslow, N. E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models, *J. Amer. Statist. Assoc.*, **88**, 9-25.
- Carlin, B.P. and Gelfand, A. E. (1990) Approaches for empirical Bayesian confidence intervals. *J. Amer. Statist. Assoc.*, **84**, 717-726.
- Castillo, J. and Lee, Y. (2007). GLM method for volatility models. to appear in *Statistical Modelling*.
- Chaganty, N. R. and Joe, H. (2006) Range of correlation matrices for dependent Bernoulli random variables. *Biometrika*, **93**, 197-206.
- Cox, D. R. (1958). The interpretation of the effects of non-additivity in the Latin square. *Biometrika*, **45**, 69-73.
- Crowder, M. J. (1995) On the use a working correlation matrix in using generalised linear models for repeated measures. *Biometrika*. **82**, 407-10.
- Davison A.C. (1986) Approximate predictive likelihood, *Biometrika*, **73**, 323-332.
- Diggle, P. J., Liang, K. Y. and Zeger, S. L. (1996) *Analysis of longitudinal data*. Oxford Univ. Press, New York.
- Drum, M.L. and McCullagh, P. (1993). REML estimation with exact covariance in the logistic mixed model, *Biometrics*, **49**, 677-689.
- Efron, B. (2003) A conversation with good friends. *Statist. Sci.*, **18**, 268-281.
- Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statist. Sci.*, **11**, 89-121.
- Fisher, R. A. (1921) On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, **1**, 3-32.

- 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 approaches for frailty models, *Biometrika*, **92**, 717-723.
- Ha, I. D., Lee, Y. and Song, J.-K. (2002). Hierarchical likelihood approach for mixed linear models with censored data. *Lifetime Data Analysis*, **8**, 163-176.
- Heagerty, P. J. and Zeger, S. (2000) Marginalized multilevel models and likelihood inference (with discussion). *Statist. Sci.*, **15**, 1-26.
- Jang, M., Lee, Y., Lawson, A., Browne, W. (2007). A comparison of the hierarchical likelihood and Bayesian approaches to spatial epidemiological modelling. *Environmetrics*, **18**, 809-821.
- Jansen, I., Beunckens, C., Molenberghs, G., Verberke, G., and Mallinckrodt, C. (2006). Analyzing incomplete discrete longitudinal clinical trial data. *Statist. Sci.*, **21**, 52-69.
- Joe, H. (1997) *Multivariate Models and Dependence Concepts*. Chapman and Hall, London.
- Kass, R.E. and Steffey, D. (1989) Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *J. Amer. Statist. Assoc.*, **84**, 717-726.
- Laird N. M. and Louis, T. A. (1987). Empirical Bayes confidence intervals based on Bootstrap samples. *J. Amer. Statist. Assoc.*, **82**, 739-750.
- Lee, Y. and Ha, I.D..(2008). H-likelihood versus orthodox BLUP in Tweedie mixed models. A paper prepared for submission.
- Lee, Y., Jang, M. and Lee, W. (2007). Hierarchical likelihood approach to standard errors of prediction in disease mapping. A paper prepared for submission.
- 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. (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. *Statistical Modelling*, **1**, 3-16.
- Lee, Y. and Nelder, J. A. (2002). Analysis of the ulcer data using hierarchical generalised linear models. *Statist. Med.*, **21**, 191-202.
- Lee, Y. and Nelder, J. A. (2004). Conditional and marginal models: another view (with discussion). *Statist. Sci.*, **19**, 219-238.
- Lee, Y. and Nelder, J. A. (2005). Likelihood for random-effects (with discussion). *Statistical and Operational Research Transactions.*, **29**, 141-182.
- Lee, Y. and Nelder, J. A. (2006a) Double hierarchical generalized linear models (with discussion). *Applied Statist.*, **55**, 139-185.
- Lee, Y. and Nelder, J. A. (2006b). Fitting via alternative random-effect models, *Statist. Comp.*, **16**, 69-75.
- Lee, Y, Nelder, J. A and Pawitan, Y. (2006) *Generalized Linear Models with Random Effects: Unified Analysis via H-Likelihood*. Chapman and Hall, London.
- Leroux, B. G., Lin, X., and Breslow, N. (1999). Estimation of disease rates in small areas: A new mixed model for spatial dependence. In *Statistical Models in Epidemiology, the Environment and Clinical Trials*, Halloran, M. E. and Berry, D. (eds.), Springer-Verlag, New York, 135-178.
- Lin, X. and Breslow, N.E. (1996). Bias correction in generalised linear mixed models with multiple components of dispersion, *J. Amer. Statist. Assoc.*, **91**, 1007-1016.

- Lindsey, J. K. and Lambert, P. (1998). On the appropriateness of marginal models for repeated measurements in clinical trials. *Statist. Med.* **17**, 447-469.
- Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley, New York.
- Ma, R. and Jorgensen, B. (2007) Nested generalized linear mixed models: Orthodox best linear unbiased predictor approach. *J. R. Statist. Soc. B*, **69**, 625-641.
- MacNab, Y. C., Farrell, P. J., Gustafson, P., and Wen. S. (2004). Estimation in Bayesian disease mapping. *Biometrics*, **60**, 865-873.
- Mathiasen P.E. (1979). Predictive function. *Scandinavian Journal of Statistics*, **6**, 1-21.
- McCullagh P. and Nelder, J. A. (1989) *Generalized Linear Models*. 2nd edn. Chapman and Hall, London.
- Molenberghs, G., Beunckens, C., Sotito, C., and Kenward, M.G. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *J. R. Statist. Soc. B.* **70**, 371-388.
- Molenberghs, G. and Lesaffre E. (1994). Marginal modelling of correlated ordinal data using a multivariate Plackett distribution. *J. Amer. Statist. Assoc.*, **89**, 633-644.
- Molenberghs, G., Verbeke, G. and Demétrio, C. (2007). An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime Data Analysis*, **13**, 513-531.
- Morris, C. N. (1983). Parametric empirical Bayes inference: theory and application, *J. Amer. Statist. Assoc.*, **78**, 47-59.
- Nelder, J.A. (1954). The interpretation of negative components of variance. *Biometrika*, **41**, 544-548.

- Nelder, J. A. (1994) The statistics of linear models: back to basics.
Statist. Comp., **4**, 221-234.
- Neyman, J. (1935) Statistical problems in agricultural experimentation
(with discussion). *J. R. Statist. Soc. B*, **2** (suppl), 107-108.
- Noh, M and Lee, Y. (2007a). Robust modelling for inference from GLM
classes. *J. Amer. Statist. Assoc.*, **102**, 1059-1072.
- Noh, M and Lee, Y. (2007b). REML estimation for binary data in GLMMs.
J. Multi. Anal, **98**, 896-915.
- Noh, M and Lee, Y. (2008). Hierarchical-likelihood approach for nonlinear
mixed-effects models. *Comp. Stat. Data. Anal.* **52**, 3517-3527..
- Noh, M., Pawitan, Y. and Lee, Y. (2005). Robust ascertainment-adjusted
parameter estimation. *Gen. Epidem.*, **29**, 68-75.
- Pearson, K. (1920) The fundamental problems of practical statistics.
Biometrika, **13**, 1-16.
- Rasch, G. (1960) *Probabilistic Models for Some Intelligence and Attainment
Tests*. Danmarks Padagogiske Institute, Copenhagen.
- Robinson, G. K. (1991) That BLUP is a good thing: The estimation of
random effects, *Statist. Sci.*, **6**, 15-51.
- Rubin, D. B. (2005). Causal inference using potential outcomes: design,
modelling, decisions. *J. Amer. Statist. Assoc.*, **88**, 9-25.
- Rubin, D. B. (2006). Causal inference through potential outcomes and
principal stratification: application to studies with "censoring" due to
death (with discussion). *Statist. Sci.*, **21**, 299-312.
- Schall, R. (1991). Estimation in generalized linear models with random
effects, *Biometrika*, **78**, 719-727.
- Shun, Z. (1997). Another look at the salamander mating data: a modified
Laplace approximation approach, *J. Amer. Statist. Assoc.*, **92**, 341-349.

- Shun, Z. and McCullagh, P. (1995). Laplace approximation of high-dimensional integrals, *J. R. Statist. Soc., B*, **57**, 749-760.
- Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman and Hall, London.
- Skrondal, A. and Rabe-Hesketh, S. (2007). Latent variable modeling: a survey. *Scan. J. Statist*, in press.
- Thomas, A., O'Hara, B., Ligges, U. and Sturtz, S. (2006). Making BUGS open. *R News*, **6**, 12-16.
- Vaida, F. and Meng X. L. (2004), Mixed linear models and the EM algorithm, in *Applied Bayesian and Causal Inference from an Incomplete Data Perspective*, edited by Gelman, A. and Meng, X. L., Wiley, New York..
- Wilk, M. B. and Kempthorne, O. (1957). Standard errors of prediction in generalized linear mixed models. *J. Amer. Statist. Assoc.*, **52**, 218-236.
- Yun, S. and Lee, Y. (2006). Robust estimation in mixed linear models with non-monotone missingness. *Statist. Med.*, **25**, 3877-3892.
- Yun, S., Lee, Y. and Kenward, M. (2007). Using Hierarchical Likelihood for Missing Data Problems. To appear in the December issue of *Biometrika*.
- Zeger, S. L., Liang, K. Y. and Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, **44**, 1049-1060.
- Zhao, Y., Staudenmayer, J., Coull, B. A. and Wand, M. P. (2006). General design Bayesian generalized linear models. *Statist. Sci.*, **21**, 35-51.