Statistics and Computing **12**: 163–174, 2002

A general maximum likelihood analysis of measurement error in generalized linear models

MURRAY AITKIN* and ROBERTO ROCCI†

*Department of Statistics, University of Newcastle, UK
†Department SEGeS, University of Molise, Campobasso, Italy

This paper describes an EM algorithm for maximum likelihood estimation in generalized linear models (GLMs) with continuous measurement error in the explanatory variables. The algorithm is an adaptation of that for nonparametric maximum likelihood (NPML) estimation in overdispersed GLMs described in Aitkin (Statistics and Computing 6: 251–262, 1996). The measurement error distribution can be of any specified form, though the implementation described assumes normal measurement error. Neither the reliability nor the distribution of the true score of the variables with measurement error has to be known, nor are instrumental variables or replication required.

Standard errors can be obtained by omitting individual variables from the model, as in Aitkin (1996).

Several examples are given, of normal and Bernoulli response variables.

Keywords: measurement error, random effects GLM, EM algorithm, mixture model, Gaussian quadrature, nonparametric maximum likelihood

1. Introduction

Measurement error is a pervasive problem in many areas of social research. There is universal agreement on the importance of allowing for measurement error, and on the potentially serious consequences of failing to do so. The practical implementation of procedures to make this allowance however lags far behind the need, as a perusal of the main references in the area (Fuller 1987, Carroll, Ruppert and Stefanski 1995) makes clear. The latter book gives extensive references to recent work: there has been a dramatic growth in the literature of the field in the last 10 years.

The simplest example of the nature of the difficulty is the normal regression model with a single unreliable explanatory variable. Adopting the notation and "error calibration model" of Carroll, Ruppert and Stefanski (1995), let Y and W be the observed response and explanatory variables, and X be the unobserved true score corresponding to the observed W. A common model—the "non-differential measurement error model"—for Y and W is that these are conditionally independent given X, with

$$Y \mid X, W \sim N(\alpha + \beta X, \sigma^2),\ W \mid X \sim N\left(X, \sigma_u^2\right).$$

Thus the measurement error in W is normally distributed with error variance σ_u^2.

Since X is unobserved, the model likelihood cannot be written down without a model assumption for the true score X. The usual assumption of a normal $X : X \sim N(\mu_X, \sigma_X^2)$ gives a six-parameter model, but only five sufficient statistics from the joint normal distribution of (Y, W). The model is not identified, and some assumption about one of the parameters must be made to allow ML estimation. Thus the intercept α must be known, or the true-score variance, or the measurement error variance (or an independent estimate of it), or the reliability of W. This is a serious difficulty as it means that internal or external reliability assessment of the covariate W is essential, and it must be assumed in addition that any externally assessed reliability applies to the particular study being analysed. Much of the discussion in Carroll, Ruppert and Stefanski deals with internal replication or instrumental variable methods for assessing the measurement error variance or the reliability. A surprising result is that if X is *non*-normal, the model *is* identified (Reiersøl 1950), and these difficulties disappear! This result is an important aid in the analysis given below.

Even greater difficulties attend non-linear models. For example, if Y is Bernoulli and has a logistic regression on the true

score X, while the observed W is normally distributed about X as above, the likelihood cannot be written analytically even with the normal assumption for X, as the normal distribution is not conjugate to the logistic likelihood. Schafer (1987) considered the generalized linear model with normal measurement error, and evaluated several approximations to the exact likelihood to avoid the Gaussian quadrature used by Hinde (1982) for the closely related overdispersion model; all the approximate methods required an independent estimate of the measurement error covariance matrix. Nakamura (1990) and Hanfelt and Liang (1997) required the stronger condition of a *known* (up to a scalar multiplier) measurement error covariance matrix.

These difficulties of maximum likelihood estimation in measurement error models are well-known and have proved resistant to treatment: most of the methods discussed by Fuller and by Carroll, Ruppert and Stefanski use other methods of estimation (see for example Carroll and Stefanski 1990 and Gleser 1990). Recent work on nonparametric maximum likelihood (NPML) estimation (Carroll, Ruppert and Stefanski 1995 §14.3, Aitkin 1996, 1999, Roeder, Carroll and Lindsay 1996) however provides a solution to these difficulties of an unexpected kind. For models which are identifiable with normal true-score distributions, Gaussian quadrature may be carried out efficiently using finite mixture maximum likelihood methods (Hinde and Wood 1987, Aitkin 1996, 1999), and general true-score distributions can be handled by estimating this distribution as part of the model, using the same methods used for other random-effect models in which the random effect distribution is estimated by NPML. Further, there is an important gain in *robustness* against incorrect parametric model assumptions for the distribution of the true score X, this has been demonstrated for the closely related variance component model by Heckman and Singer (1984) and Davies (1987) (see Aitkin 1996 for discussion).

It is generally considered that replication or known true scores are needed for at least part of the data for the normal error score model to estimate the measurement error variance: we show that this is not necessary except in the normal true-score model which is otherwise unidentifiable. We illustrate the effect of the magnitude of the measurement error on the other model parameters by computing the profile likelihood and the other model parameters and their standard errors for a grid of values of the measurement error variance. This also provides a full examination of the sensitivity of model conclusions to the size of the measurement error.

This paper gives the general theory of this approach for measurement error in generalized linear models, and illustrates it with applications to normal regression models. Applications to other exponential family models will be discussed elsewhere. Examples are given of these cases using GLIM4 implementations of an EM algorithm adapted from the EM algorithm of Aitkin and Francis (1995).

The computational method described here is very general, since it allows an arbitrary true-score distribution, any exponential family response distribution, any specified measurement error model, and additional covariates measured without error. The extension to multiple error-prone covariates is straightforward, and involves little additional computation.

We begin for simplicity of exposition with the normal regression model.

2. The normal linear model with a single error-prone covariate

We begin with a more general version of the model above, and assume that the true-score distribution is normal. Suppose that we have in addition to W a set (vector) of error-free covariates Z, and the model for $Y \mid W, X, Z$ is

$$Y \mid W, X, Z \sim N(\alpha + \beta X + \gamma' Z, \sigma^2),\ W \mid X, Z \sim N\big(X, \sigma_u^2\big),$$
$$X \mid Z \sim N\big(\mu + \lambda' Z, \sigma_x^2\big).$$

We allow the true score X to depend on the error-free covariates Z through a regression. A distribution is not needed for Z, as it is fully observed, and we now suppress notationally the conditioning on Z.

The joint marginal distribution $h(Y, W \mid \theta)$ of the observables Y and W given Z and all the parameters $\theta = (\alpha, \beta, \gamma, \mu, \lambda, \sigma^2, \sigma_u^2, \sigma_x^2)$ is given by

$$h(Y, W \mid \theta) = \int f(Y \mid W, X, \theta)\, m(W \mid X, \theta)\, \pi(X \mid \theta)\, dX$$

where f is the response model density, m the measurement model density and π the true-score density. The marginal density h is a mixture over X with respect to its marginal density π.

For subsequent analyses we now transform this model so that the transformed true-score $X^* = X - \lambda' Z$ has a homogeneous $N(\mu, \sigma_x^2)$ distribution; then the original true-score can be expressed as $X = X^* + \lambda' Z$, and defining $\gamma^* = \gamma + \beta\lambda$ and dropping the stars, the above model transforms to

$$Y \mid W, X \sim N(\alpha + \beta X + \gamma' Z, \sigma^2),\ W \mid X \sim N\big(X + \lambda' Z, \sigma_u^2\big),$$
$$X \sim N\big(\mu, \sigma_x^2\big).$$

Thus the regression of the true score X on Z can be transformed to that of the observed score W on Z, with the same regression coefficient. (A further generalisation of the original model might be considered, in which the observed score W *and* the true score X *both* depend on the covariate Z, but it is easily shown that these two regression coefficients are not separately identifiable.)

The reliability of W is $\rho = \sigma_x^2/(\sigma_u^2 + \sigma_x^2)$. The joint marginal distribution of the observables (Y, W) is conveniently expressed as

$$\begin{bmatrix} Y \\ W \end{bmatrix} \sim N\left(\begin{bmatrix} \alpha + \beta\mu + \gamma' Z \\ \mu + \lambda' Z \end{bmatrix}, \begin{bmatrix} \sigma^2 + \beta^2\sigma_x^2 & \beta\sigma_x^2 \\ \beta\sigma_x^2 & \sigma_x^2 + \sigma_u^2 \end{bmatrix} \right).$$

This model is unidentifiable; we identify it by fixing σ_u^2 and estimating the other parameters as functions of σ_u^2; this will be used

in other identifiable models to construct a profile likelihood in σ_u^2 and to establish the sensitivity of the other parameter estimates to the magnitude of the measurement error.

Since σ_x^2 can be estimated from the data as $\sigma_w^2 - \sigma_u^2$, the reliability can be estimated, and consistent estimates obtained by correcting the estimated regression coefficients of Y on W and Z. However these estimates are not efficient as additional information is available about β in the variance of Y and the covariance of Y and W.

We estimate the model parameters by maximum likelihood using an EM algorithm (Dempster, Laird and Rubin 1977).

3. Maximum likelihood by EM

We treat the true scores as missing data. Given the observations (y_i, w_i, X_i, z_i), $i = 1, \ldots, n$, the complete-data likelihood is

$$L^* = \prod_{i=1}^{n} f(y_i \mid X_i) m(w_i \mid X_i) \pi(X_i),$$

and the corresponding log-likelihood is

$$\begin{aligned}\ell^* = \sum_{i=1}^{n} \Big\{ &-\frac{1}{2}\log(2\pi) - \log\sigma - \frac{1}{2\sigma^2}[y_i - \alpha - \beta X_i - \gamma' z_i]^2 \\ &-\frac{1}{2}\log(2\pi) - \log\sigma_u - \frac{1}{2\sigma_u^2}[w_i - X_i - \lambda' z_i]^2 \\ &-\frac{1}{2}\log(2\pi) - \log\sigma_x - \frac{1}{2\sigma_x^2}[X_i - \mu]^2 \Big\}.\end{aligned}$$

The missing data appear in the complete data log-likelihood as X_i and X_i^2: these are replaced in the E step by their conditional expectations $\tilde{x}_i$ and $\tilde{x}_i^2 + v_i$ given the observed data (y_i, w_i, z_i), where v_i is the conditional variance of X_i.

Thus in the M step the ML estimates of the parameters are the solutions of

$$\begin{aligned}\sum_i y_i &= n\alpha + \beta \sum_i \tilde{x}_i + \gamma' \sum z_i \\ \sum_i \tilde{x}_i y_i &= \alpha \sum_i \tilde{x}_i + \beta \sum_i \left(\tilde{x}_i^2 + v_i\right) + \gamma' \sum \tilde{x}_i z_i \\ \sum_i z_i y_i &= \alpha \sum_i z_i + \beta \sum_i \tilde{x}_i z_i + \gamma' \sum z_i z_i' \\ \sigma^2 &= \sum_i [(y_i - \alpha - \beta\tilde{x}_i - \gamma' z_i)^2 + \beta^2 v_i] \Big/ n \\ \sum_i z_i (w_i - \tilde{x}_i) &= \lambda' \sum_i z_i z_i' \\ \mu &= \sum_i \tilde{x}_i \Big/ n \\ \sigma_x^2 &= \sum_i [(\tilde{x}_i - \mu)^2 + v_i] \Big/ n\end{aligned}$$

The regression parameter estimates for the response model are ridge estimates, with a loading on the diagonal of the SSP matrix of 0 for α and γ, and $\sum_i v_i$ for β. The regression parameter estimates for the measurement model are uncentred regression estimates with an offset of $\tilde{x}_i$ and no intercept term. The measurement error variance σ_u^2 is fixed.

For the E step, the conditional distribution of X given y_i, w_i and z_i is normal, with mean (after considerable algebra)

$$\tilde{x}_i = \mu + \phi e_{y_i \cdot \mu} \big/ \beta + \rho(1-\phi)\, e_{w_i \cdot \mu}$$

and (constant) variance

$$v_i = v = (1-\rho)(1-\phi)\sigma_x^2,$$

where

$$\begin{aligned}e_{y_i \cdot \mu} &= y_i - \alpha - \beta\mu - \gamma' z_i, \\ e_{w_i \cdot \mu} &= w_i - \mu - \lambda' z_i, \\ \phi &= \frac{\rho(1-\rho)\beta^2\sigma_w^2}{\sigma^2 + \rho(1-\rho)\beta^2\sigma_w^2} \\ &= \frac{(1-\rho)\beta^2\sigma_x^2}{\sigma^2 + (1-\rho)\beta^2\sigma_x^2}.\end{aligned}$$

The expectation is more conveniently computed recursively: write at the r-th iteration

$$\begin{aligned}e^{(r)}_{y_i \cdot \tilde{x}_i} &= y_i - \alpha - \beta\tilde{x}_i^{(r)} - \gamma' z_i, \\ e^{(r)}_{w_i \cdot \tilde{x}_i} &= w_i - \tilde{x}_i^{(r)} - \lambda' z_i;\end{aligned}$$

these quantities are the model residuals at the current $\tilde{x}_i$. Then $\tilde{x}_i$ can be computed equivalently as

$$\begin{aligned}\tilde{x}_i^{(r+1)} = \tilde{x}_i^{(r)} &+ \phi e^{(r)}_{y_i \cdot \tilde{x}_i} \big/ \beta + \rho(1-\phi) e^{(r)}_{w_i \cdot \tilde{x}_i} \\ &- (1-\rho)(1-\phi)\big(\tilde{x}_i^{(r)} - \mu\big).\end{aligned}$$

The EM algorithm alternates between E and M steps; the initial M step is conveniently taken as the regression of Y on W and Z. The maximized likelihood is computed in each iteration from the joint distribution of (Y, W) (all constants are included in the likelihood calculation). A GLIM4 implementation of this algorithm is available from the authors.

4. Example

We analyse the SOLV example from Aitkin *et al.* (1989) Chapter 2. A sample of twenty-four children was randomly drawn from the population of fifth-grade children attending a state primary school in a Sydney suburb. Each child was assigned to one of two experimental groups, and given instructions by the experimenter on how to construct, from nine differently coloured blocks, one of the 3×3 square designs in the Block Design subtest of the Wechsler Intelligence Scale for Children (WISC). Children in the first group were told to construct the design by starting with a row of three blocks (row group), and those in the second group were told to start with a corner of three

blocks (corner group). The total time in seconds to construct four different designs was then measured for each child.

Before the experiment began, the extent of each child's "field dependence" was tested by the Embedded Figures Test, which measures the extent to which subjects can abstract the essential logical structure of a problem from its context (high scores corresponding to high field dependence and low ability).

The data are given below. The experimenter was interested in whether the different instructions produced any change in the average time required to construct the designs, and whether this time was affected by field dependence. The model fitted is a normal regression of TIME on EFT and treatment GROUP dummy.

```
Row GROUP
TIME: 317 464 525 298 491 196 268 372 370  739  430 410
EFT    59  33  49  69  65  26  29  62  31  139   74  31
Corner GROUP
TIME: 342 222 219 513 295 285 408 543 298  494  317 407
EFT:   48  23   9 128  44  49  87  43  55   58  113   7
```

The reliability of the Embedded Figures Test is unknown. We examine the effect of allowing for its unreliability. We fit the model over a grid of values of the measurement error variance; the maximized (profile) likelihood is constant over this grid, as the full model is unidentifiable. Parameter values and the standard error for the treatment effect (computed as in Aitkin 1996, 1999) are given in Table 1. Here W is the EFT score and Z the treatment group dummy variable; ρ is the reliability of EFT.

The intercept and the regression coefficient of time on EFT true-score are strongly dependent on the measurement error variance, as expected, but the treatment estimate and its standard error are almost unaffected by measurement error. The near-invariance of these estimates is due to the near-independence of EFT and the treatment group, a consequence of the randomized assignment of children to group.

It should be noted that the GROUP estimate for the "perfectly reliable" covariate with $\sigma_u^2 = 0$, namely $\hat{\gamma} = -44.75$, is not quite equal to that for the usual normal regression model (-44.24), since the former is corrected for regression of the covariate on the GROUP factor. The GROUP difference on EFT is -0.25; correcting the GROUP effect on TIME for this difference gives the adjusted effect of $-44.24 + (2.038 \times -0.25) = -44.75$.

Table 1. *Parameter estimates for SOLV data*

σ_u^2	ρ	α	β	γ	SE(γ)	σ	$-2\log L$
0	1.000	293.4	2.038	−44.75	42.28	101.1	525.96
100	0.909	282.1	2.241	−44.75	42.28	98.9	525.96
200	0.819	268.3	2.489	−44.75	42.28	96.0	525.96
300	0.728	251.1	2.799	−44.75	42.28	92.3	525.96
400	0.637	229.0	3.197	−44.75	42.28	87.3	525.96
500	0.547	199.4	3.728	−44.75	42.10	80.3	525.95
600	0.455	157.0	4.492	−44.75	42.47	69.5	525.95

Since the profile likelihood is effectively flat in σ_u^2 for both the GROUP model and the null model of no group difference, the likelihood ratio test for the GROUP effect ($\gamma = 0$) is unaffected by the measurement error in EFT. The deviance difference is 1.11, clearly not significant.

We now extend this approach to discrete true-score distributions.

5. Discrete true-score distributions for quadrature

In generalized linear (non-normal) response models, the normal true-score model leads to an *identifiable* mixture, but also to a non-analytic likelihood, as noted in §1. We attack this problem by direct numerical integration of the likelihood over the true-score distribution, by replacing the integral by a finite sum over Gaussian quadrature mass-points X_k, with masses π_k. This approach is described in detail in Aitkin (1996) for the closely related overdispersion model with a normal "extra variation" term.

We first re-parametrise the model further so that the true-score distribution is standardised. Define

$$X_s = (X - \mu)/\sigma_x$$

so that

$$X = \mu + \sigma_x X_s.$$

Then the response model linear predictor becomes

$$\alpha + \beta(\mu + \sigma_x X_s + \lambda' z) + \gamma' z = \alpha^* + \beta^* X_s + \gamma^{*'} z$$

where

$$\alpha^* = \alpha + \beta\mu, \quad \beta^* = \beta\sigma_x, \quad \gamma^* = \gamma + \beta\lambda,$$

and that for the measurement model becomes

$$\mu + \sigma_x X_s + \lambda' z.$$

Given the observations (y_i, w_i, z_i), $i = 1, \ldots, n$, the likelihood is

$$L(\theta) = \prod_{i=1}^{n} \int f(y_i \mid X_s, \theta) m(w_i \mid X_s, \theta) \pi(X_s)\, dX_s.$$

We approximate the integral by the sum:

$$L(\theta) \doteq \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f(y_i \mid X_k, \theta) m(w_i \mid X_k, \theta)$$

where the X_k and π_k are the Gaussian quadrature mass-points and masses, which are tabulated for each K in standard sources, e.g. Abramowitz and Stegun (1964).

The maximization of the likelihood over θ is now (approximately) equivalent to a finite mixture maximum likelihood

problem, in which we maximize over θ the mixture likelihood

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \left(\sum_{k=1}^{K} \pi_k f(y_i \mid X_k, \theta) m(w_i \mid X_k, \theta) \right) \\ &= \prod_i \left(\sum_k \pi_k f_{ik} m_{ik} \right) \\ &= \prod_i \left(\sum_k \pi_k h_{ik} \right) \\ &= \prod_i h_i . \end{aligned}$$

A very closely related mixture ML problem is extensively discussed in Aitkin (1996); it differs from that above only in the omission of the measurement model term m_{ik}. The EM algorithm described there for overdispersed GLMs is readily adapted to the measurement error model above. It should be noted that the normal distribution m of measurement error is not a critical assumption of the model—any specified measurement error distribution can be used, with a corresponding change to the mass-points, masses and likelihood in the following argument.

6. Finite mixture maximum likelihood

To maximize the likelihood, we proceed as in Aitkin (1996). The mixture log-likelihood can be written

$$\ell = \sum_i \log \left(\sum_k \pi_k h_{ik} \right)$$

and

$$\begin{aligned} \frac{\partial \ell}{\partial \theta} &= \sum_i \left(\sum_k \pi_k h_{ik} \frac{\partial \log h_{ik}}{\partial \theta} \right) \Big/ h_i \\ &= \sum_i \sum_k w_{ik} \frac{\partial \log h_{ik}}{\partial \theta} \end{aligned}$$

where the weights w_{ik} are given by

$$w_{ik} = \frac{\pi_k f_{ik} m_{ik}}{\sum_\ell \pi_\ell f_{i\ell} m_{i\ell}} .$$

Thus maximizing the likelihood with respect to θ is equivalent to solving a weighted ML problem for the joint response/measurement model h. Alternately maximizing the likelihood for fixed weights, and re-estimating the weights at the new parameter estimates, is an EM algorithm (Dempster, Laird and Rubin 1977, Aitkin 1996). The response model f and the measurement model m will usually depend on functionally distinct parameters and so the M-step can be split into two parts, for f and m separately, as for the normal example with known measurement error above.

A notable difference from the earlier model in the score equations is that the X_k are now *known*, and are not being "imputed" in the E-step; what is being imputed is the *weights*—the posterior probabilities that each observation i comes from the *known* true score X_k. Thus no adjustment of the SSP matrix is required—all the M-step calculations are exact ML with weights w_{ik}. The scaling of the true score to be standardised also changes the measurement model: the true-score X_s now has an unknown regression coefficient, the true-score standard deviation. Thus there is no offset in the measurement model: both models estimate all the parameters except the measurement error variance. After convergence to the MLEs of the parameters, the MLEs of α, β and γ are obtained by rescaling.

We illustrate with the SOLV example, using 20-point quadrature. The model is as before formally unidentifiable for normal X, and we take σ_u^2 to be known. Parameter estimates for α, β, γ and σ are given in Table 2, over the same grid of values of σ_u^2. The profile likelihood in σ_u^2 is almost flat, as expected, over the range tabulated. The variations in the likelihood are due to the approximation of the exact normal model by the 20-point discrete distribution. Parameter estimates from the numerical integration are very close to those from the analytic model above; the departures increase with increasing error variance.

For small values (near zero) of σ_u^2 the deviance changes because the normal density representation for W starts to break down for very small variances. This is illustrated in Table 3, in which the grid of σ_u^2 is extended down to 10. This is of little practical consequence because the extreme corresponds to a reliability near 1.00. However if a small number of quadrature points is used, real variations in the deviance can occur over the range of σ_u^2, because a non-normal discrete distribution is being used, and this is formally identifiable. We illustrate this effect in Table 4, for fixed $\sigma_u^2 = 100$. It is notable that quadrature with $K = 5$–10 mass-points gives a lower deviance than for $K = 20$:

Table 2. *Parameter estimates for SOLV data by GQ*

σ_u^2	α	β	γ	σ	$-2 \log L$
100	283.0	2.225	−44.19	99.2	526.07
200	268.2	2.491	−44.13	96.0	526.00
300	251.4	2.793	−44.05	92.4	525.97
400	230.1	3.177	−43.96	87.7	525.96
500	201.9	3.685	−43.83	81.1	525.96
600	164.8	4.352	−43.66	71.5	525.97

Table 3. *Parameter estimates for SOLV data by GQ*

σ_u^2	ρ	α	β	γ	σ	$-2 \log L$
10	0.990	294.1	2.026	−44.24	101.6	523.91
20	0.980	291.5	2.072	−44.23	100.8	523.67
30	0.971	291.0	2.080	−44.23	100.8	524.51
40	0.962	290.4	2.093	−44.23	100.7	525.14
50	0.953	289.6	2.105	−44.22	100.6	525.60
60	0.944	288.9	2.119	−44.22	100.5	525.96
70	0.935	288.2	2.132	−44.22	100.5	526.25
80	0.926	287.4	2.146	−44.21	100.4	526.50
90	0.918	286.6	2.160	−44.21	100.3	526.72

Table 4. *Parameter estimates for SOLV data by GQ*

K	α	β	γ	σ	$-2\log L$
2	296.9	1.975	−44.26	108.3	555.50
3	290.7	2.086	−44.23	102.8	527.54
4	295.0	2.009	−44.25	103.8	526.93
5	276.5	2.343	−44.16	95.5	518.26
6	285.9	2.172	−44.21	99.3	519.52
7	278.7	2.302	−44.17	96.4	516.15
8	287.9	2.137	−44.22	100.0	517.71
9	282.2	2.239	−44.19	97.8	516.49
10	286.8	2.120	−44.22	100.4	517.93
20	283.0	2.225	−44.19	99.2	526.07

the "strongly discrete" normal model gives a better "fit" to the data than the "fine mesh" normal model. This is a consequence of the actual (unknown) distribution of the true scores: all the models have the same number of parameters. Despite this variation, the treatment effect is almost invariant to the size of the measurement error variance or the number of mass-points, a consequence of the orthogonality of the design.

We now extend this approach to arbitrary (non-normal) true-score distributions.

7. Estimating the true-score distribution

We consider the model as before, but now make no assumption about $\pi(X)$, the true-score distribution, other than that it is *non*-normal, and therefore at least formally identifiable. Given the observations (y_i, w_i, z_i), $i = 1, \ldots, n$, the likelihood is

$$L(\theta, \pi) = \prod_{i=1}^{n} \int f(y_i \mid X, \theta) m(w_i \mid X, \theta) \pi(X)\, dX.$$

We regard the distribution $\pi(X)$ as part of the unknown structure to be estimated. Invoking the results of Kiefer and Wolfowitz (1956), Laird (1978) and Lindsay (1983), we know that the NPML estimate of π is a discrete distribution on a finite number of mass-points. The number K, locations X_k and masses π_k of these mass-points have to be determined as parameters of the model. Thus the maximization of the likelihood over θ and π is equivalent to a finite mixture maximum likelihood problem, in which we maximize the mixture likelihood

$$\begin{aligned} L &= \prod_{i=1}^{n} \left(\sum_{k=1}^{K} \pi_k f(y_i \mid X_k) m(w_i \mid X_k) \right) \\ &= \prod_i \left(\sum_k \pi_k f_{ik} m_{ik} \right) \\ &= \prod_i \left(\sum_k \pi_k h_{ik} \right) \\ &= \prod_i h_i \end{aligned}$$

over θ, K, X_k and π_k.

This mixture ML problem is also extensively discussed in Aitkin (1996); it differs from that above only in the omission of the measurement model term m_{ik}. The EM algorithm described there for overdispersed GLMs is readily adapted to the measurement error model above. The normal distribution m of measurement error is again not a critical assumption of the model—any specified measurement error distribution can be used, with a corresponding change to the likelihood in the following argument.

8. Finite mixture maximum likelihood

To maximize the likelihood, we proceed as in §6, with the additional estimation of the π_k and X_k. The mixture log-likelihood can be written as before

$$\ell = \sum_i \log \left(\sum_k \pi_k h_{ik} \right)$$

and

$$\begin{aligned} \frac{\partial \ell}{\partial \theta} &= \sum_i \left(\sum_k \pi_k h_{ik} \frac{\partial \log h_{ik}}{\partial \theta} \right) \Big/ h_i \\ &= \sum_i \sum_k w_{ik} \frac{\partial \log h_{ik}}{\partial \theta} \end{aligned}$$

where the weights w_{ik} are as before

$$w_{ik} = \frac{\pi_k f_{ik} m_{ik}}{\sum_\ell \pi_\ell f_{i\ell} m_{i\ell}}.$$

The mixture proportions π_k are easily shown to be estimated by $\hat{\pi}_k = \sum_i w_{ik}/n$ as in Aitkin (1996). It remains to estimate the true-score mass-point locations X_k. In the general GLM these have to be obtained by Newton methods, but in the normal model with normal measurement error, $\hat{X}_k$ can be expressed explicitly as a weighted average of residuals from the two regressions and the previous estimate. As in Aitkin (1996), the number of mass-points K has to be determined sequentially, starting from 2, and is increased until the maximized likelihood stabilises.

A general question raised by a referee is the behaviour of parameter estimates in the model when the true-score distribution is itself estimated nonparametrically, since this distribution is, at least conceptually, infinite dimensional. New results by Murphy and van der Vaart (2000) show that semiparametric profile likelihoods, where the nuisance parameter profiled out may have infinite dimension, have the same asymptotic properties for maximum likelihood estimators and likelihood ratio test statistics as do profile likelihoods with finite dimensional nuisance parameters. These asymptotic properties may require larger sample sizes for validity, and so small-sample behaviour of estimators is certainly of interest. We will report these separately elsewhere.

We illustrate true-score estimation with the normal regression model.

9. The normal regression model

As above it is convenient to re-parametrise the (unknown) true-score distribution to have mean 0 and variance 1. The response and measurement log-densities are

$$\begin{aligned}\log f_{ik} &= \log f(y_i \mid X_k, \alpha^*, \beta^*, \gamma^*, \sigma)\\ &= -\frac{1}{2}\log(2\pi) - \log\sigma\\ &\quad - \frac{1}{2\sigma^2}(Y_i - \alpha^* - \beta^* X_k - \gamma^{*'} z_i)^2\\ \log m_{ik} &= \log m(w_i \mid X_k, \mu, \sigma_x, \lambda, \sigma_u)\\ &= -\frac{1}{2}\log(2\pi) - \log\sigma_u\\ &\quad - \frac{1}{2\sigma_u^2}(w_i - \mu - \sigma_x X_k - \lambda' z_i)^2.\end{aligned}$$

The likelihood equations for the parameters are simple weighted versions of the likelihood equations for the two normal models, and are not reproduced. It remains to estimate the true scores X_k. Differentiating the log-likelihood with respect to X_k, we have

$$\frac{\partial \ell}{\partial X_k} = \sum_i w_{ik}\left[\frac{\beta^*}{\sigma^2}(y_i - \alpha^* - \beta^* X_k - \gamma^{*'} z_i) + \frac{\sigma_x}{\sigma_u^2}(w_i - \mu - \sigma_x X_k - \lambda' z_i)\right].$$

The second derivative gives

$$\frac{\partial^2 \ell}{\partial X_k^2} = -\sum_i w_{ik}\left[\frac{\beta^{*2}}{\sigma^2} + \frac{\sigma_x^2}{\sigma_u^2}\right].$$

A Newton update to the current $X_k^{(r)}$ gives

$$X_k^{(r+1)} = X_k^{(r)} + \sum_i w_{ik}\left[\frac{\beta^*}{\sigma^2}e_{Y_{ik}} + \frac{\sigma_x}{\sigma_u^2}e_{W_{ik}}\right] \Big/ \sum_i w_{ik} \times \left[\frac{\beta^{*2}}{\sigma^2} + \frac{\sigma_x^2}{\sigma_u^2}\right],$$

where $e_{Y_{ik}}$ and $e_{W_{ik}}$ are the model residuals:

$$e_{Y_{ik}} = y_i - \alpha^* - \beta^* X_k - \gamma^{*'} z_i, \quad e_{W_{ik}} = w_i - \mu - \sigma_x X_k - \lambda' z_i.$$

Write

$$\phi = \frac{\beta^{*2}}{\sigma^2} \Big/ \left(\frac{\beta^{*2}}{\sigma^2} + \frac{\sigma_x^2}{\sigma_u^2}\right) = \frac{\beta^{*2}\sigma_u^2}{\sigma^2} \Big/ \left(\frac{\beta^{*2}\sigma_u^2}{\sigma^2} + \sigma_x^2\right).$$

Then the update can be expressed simply, as

$$X_k^{(r+1)} = X_k^{(r)} + \sum_i w_{ik}\left[\phi e_{Y_{ik}}/\beta^* + (1-\phi)e_{W_{ik}}/\sigma_x\right] \Big/ \sum_i w_{ik}.$$

Convergence of the EM algorithm is greatly accelerated by starting from the parameter estimates and weights from the Gaussian quadrature fit assuming a normal true-score distribution.

Note that, if true-score estimates for individual observations are required, these can be obtained as posterior means of the random effects X_i in the usual empirical Bayes framework, by weighting the estimated $\hat{X}_k$ by the estimated posterior probabilities $\hat{w}_{ik}$ that the true-score X_i takes the values $\hat{X}_k$.

Parameter values are given in Table 5 for the SOLV data with true-score distribution estimation, for $K = 3$ and $40 \le \sigma_u^2 \le 190$. The NPML estimate was either $K = 3$ or $K = 4$ for all values of σ_u^2; the four mass-point estimate was little better than the three-point estimate (in a few cases it was worse, as a local maximum was found), and the four-point estimate always contained one component with two, one or no individual observations. Table 5 gives also the estimated true-score distribution for each σ_u^2. The three-point estimate was very stable over σ_u^2 and is adopted as the best estimate below.

Again the estimate of the treatment effect is almost invariant to the measurement error variance. However an important result of the NPML analysis is that the profile likelihood in σ_u^2 is *not* flat, but has a maximum at $\sigma_u^2 = 75$, with large and very small values of σ_u^2 being very unlikely: the reliability of EFT is clearly high.

The reader may be perplexed by this result. If the normal true-score model is unidentifiable, with a flat profile likelihood in σ_u^2, how can the model be identifiable when *nothing* is known about this distribution?

We need to remember that the *only* unidentifiable model (given a normal measurement error distribution) is the normal true-score model; *any other* distribution gives an (at least formally) *identifiable* model, in which the profile likelihood in σ_u^2 will *not* be flat. So the key to this result is the distribution of the *observed* score: if the true-score is normal and the measurement error is normal, then the distribution of the residuals from the regression of the observed score W on Z *must* be normal. If the sample distribution of the W residuals is non-normal, then the true-score distribution of X *cannot* be normal, and the correct model is identifiable; without knowledge of this distribution, we use the NPML estimate, and the corresponding mixture likelihood is not flat in the measurement error variance.

Examination of the EFT residuals (the EFT values corrected for the EFT mean in each treatment group) shows that they are substantially skewed, with three large observations. A formal test for normality does not reject this hypothesis, but the residual QQ-plot departs considerably from linearity. This skewness is detected by the NPML estimation, and the true-score distribution is estimated as a three- or four-point one, depending on the value

Table 5. *Parameter estimates for SOLV data by NPML (K = 3)*

σ_u^2	α	β	γ	σ	$-2\log L$	$\hat{X}_k$, $(\hat{\pi}_k)$		
40	293.0	2.045	−44.24	102.5	515.0	1.975	0.0276	−1.020
						0.1667	0.4971	0.3363
50	293.2	2.041	−44.24	102.6	511.5	1.975	0.0285	−1.018
						0.1666	0.4962	0.3371
60	293.5	2.037	−44.24	102.7	510.0	1.975	0.0291	−1.017
						0.1666	0.4957	0.3377
70	293.6	2.034	−44.24	102.7	509.5	1.977	0.0296	−1.017
						0.1664	0.4956	0.3380
80	293.7	2.032	−44.24	102.8	509.5	1.979	0.0299	−1.016
						0.1661	0.4957	0.3381
90	293.8	2.031	−44.24	102.8	509.8	1.982	0.0303	−1.015
						0.1657	0.4961	0.3382
100	293.7	2.032	−44.24	102.9	510.3	1.987	0.0307	−1.015
						0.1650	0.4969	0.3381
110	293.6	2.034	−44.24	102.9	510.8	1.994	0.0314	−1.014
						0.1640	0.4979	0.3380
120	293.4	2.038	−44.24	102.9	511.4	2.003	0.0325	−1.013
						0.1627	0.4994	0.3379
130	293.1	2.042	−44.24	102.9	511.9	2.015	0.0335	−1.011
						0.1612	0.5011	0.3377
140	292.8	2.049	−44.24	102.8	512.5	2.030	0.0359	−1.010
						0.1590	0.5035	0.3376
150	292.2	2.060	−44.24	102.8	513.1	2.056	0.0411	−1.008
						0.1553	0.5071	0.3376
160	291.2	2.078	−44.23	102.6	513.6	2.108	0.0548	−1.005
						0.1479	0.5141	0.3380
170	290.0	2.099	−44.23	102.4	514.0	2.176	0.0699	−1.001
						0.1390	0.5222	0.3387
180	289.6	2.106	−44.22	102.4	514.4	2.183	0.0654	−0.9968
						0.1388	0.5229	0.3383
190	289.7	2.103	−44.22	102.4	514.9	2.135	0.0433	−1.004
						0.1456	0.5221	0.3323

of σ_u^2. For example, for $\sigma_u^2 = 500$, the estimated true-score distribution (on the original scale) has masses of 0.083, 0.587 and 0.330 at 157.3, 59.1 and 23.5 respectively. Posterior probabilities for each observation in each component are shown in Table 6. In inspecting these probabilities, recall that they are determined by both the TIME score and the EFT score. Component 1 is identified by the two largest observed scores, of 139 and 128, which also correspond to large values of TIME. The EFT score of 113 however does not have a high probability in this component because its TIME score is low. Component 3 is identified by the small EFT score of 9 and the score of 26; both these observations have low TIME scores.

We show in Fig. 1 the profile deviance $(-2\log L)$ plotted against the error variance σ_u^2, and in Fig. 2 the same plot against the log error variance.

The MLE of σ_u^2 is 75, and the asymptotic 95% confidence interval is (43,156). The corresponding interval for the reliability is (0.859, 0.961). The non-constant profile likelihood in σ_u^2 in the NPML analysis allows *full ML estimation of the error*

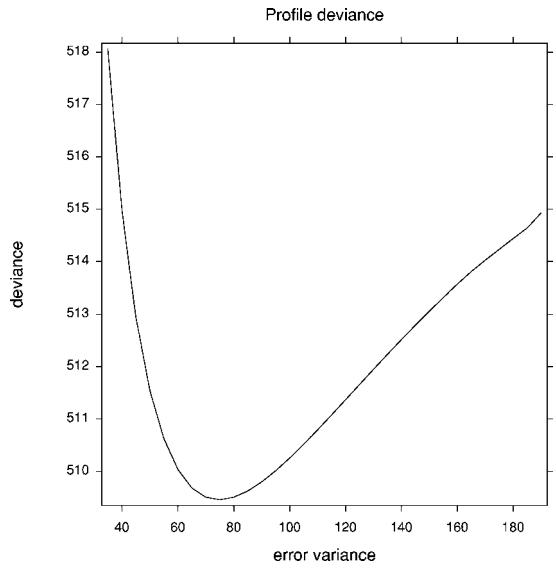

Fig. 1. *Profile deviance for measurement error variance*

Table 6. *Posterior probabilities of component membership*

i	EFT	$X_1 = 157.3$	$X_2 = 59.1$	$X_3 = 23.5$
1	59	0.000	0.672	0.328
2	33	0.000	0.730	0.270
3	49	0.000	0.927	0.073
4	69	0.000	0.738	0.262
5	65	0.000	0.953	0.047
6	26	0.000	0.082	0.918
7	29	0.000	0.192	0.808
8	62	0.000	0.818	0.182
9	31	0.000	0.456	0.544
10	139	1.000	0.000	0.000
11	74	0.000	0.942	0.057
12	31	0.000	0.569	0.431
13	48	0.000	0.711	0.289
14	23	0.000	0.143	0.857
15	9	0.000	0.071	0.929
16	128	0.973	0.027	0.000
17	44	0.000	0.538	0.462
18	49	0.000	0.576	0.424
19	87	0.003	0.974	0.023
20	43	0.000	0.948	0.052
21	55	0.000	0.685	0.315
22	58	0.001	0.959	0.041
23	113	0.083	0.975	0.016
24	7	0.000	0.365	0.635

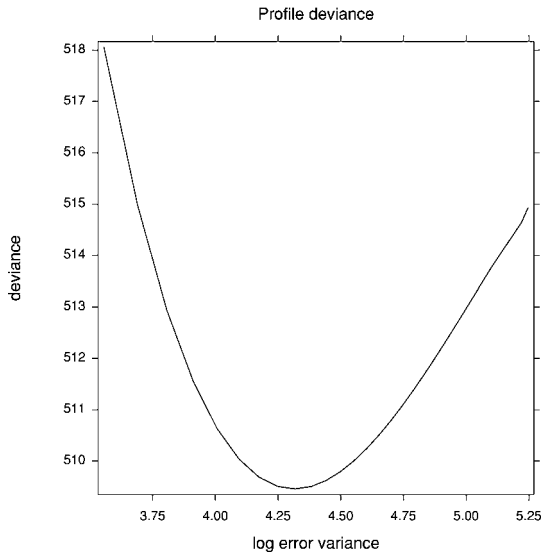

Fig. 2. *Profile deviance for measurement error log-variance*

variance, as the model is now identifiable. We give the details below.

10. Full ML estimation with unknown error variance

We now extend the NPML estimation of §7 to the case of unknown error variance. Only one change is required: in the M-step an additional score equation has to be solved for the measurement error variance:

$$\hat{\sigma}_u^2 = \sum_i \sum_k w_{ik}(w_i - \hat{\mu} - \hat{\sigma}_x X_k - \hat{\lambda}' z_i)^2 \Big/ n.$$

No other changes are necessary. Convergence is rapid from a suitable value of σ_u^2 found from the grid analysis with σ_u^2 fixed.

In comparing the NPMLE with the model ignoring measurement error ($\sigma_u = 0$), a problem occurs with the definition of the likelihood: as $\sigma_u \to 0$ the normal error distribution degenerates to a non-differentiable spike. Even small values of σ_u may give an inaccurate likelihood. For this reason we do not attempt to test for zero measurement error by direct comparison of maximized likelihoods. As in Aitkin (1996), our aim is to *allow* for the effect of measurement error, rather than *test* for it.

A related problem is how to determine whether the assumption of a normal true-score distribution is reasonable. A direct comparison of the normal model deviance with the NPML deviance is not straightforward except by bootstrapping, but we can compare by the usual likelihood ratio test the K-masspoint model with the Gaussian quadrature model with the same number of masspoints (since the constrained mixture model is not on the boundary of the parameter space). At $\hat{\sigma}_u^2 = 75$, the NPML deviance is 509.46 with $K = 3$ masspoints, compared with the Gaussian quadrature deviance of 530.89. The deviance change of 21.43 on 5 df is large; even comparing the analytic normal model deviance of 525.96 we have a deviance change of 16.50. This comparison is however complicated by the variation in the deviance with the number of masspoints by Gaussian quadrature. For example, for $\sigma_u^2 = 75$, the deviances by Gaussian quadrature with 2(1)10, 20 masspoints are 607.36, 530.89, 527.97, 516.29, 516.34, 513.74, 514.40, 513.52, 514.58, 525.62. Thus seven-point Gaussian quadrature gives a deviance only 4.28 greater than the NPML deviance with three points. It is notable that 20-point quadrature "fits" relatively badly compared with small numbers of masspoints, as noted above; this suggests that there *is* some departure from normality. Further investigation of these issues would be helpful; the bootstrap likelihood ratio test (McLachlan 1987) provides a general solution, at the cost of additional computing.

While the treatment effect is very stable in the measurement error analysis, the intercept and slope of the regression on the error-prone variable EFT are very sensitive to the measurement error variance. Thus if our interest is in *this* regression, information about the measurement error variance is critical.

We illustrate its importance with an analysis of the corn data of Fuller (1987, p.18), which are observations of corn yields and available nitrogen at 11 sites on Marshall soil in Iowa. Fuller gave an analysis based on a given measurement error variance ($\sigma_u^2 = 57$) in the available nitrogen explanatory variable, calculated from sampling considerations and measurement error in the chemical analysis. The data are shown below, ordered by nitrogen.

Yield Y	99	96	90	86	91	104	86	96	99	110	115
Nitrogen X	50	51	53	64	64	69	70	70	94	95	97
Site	7	11	3	4	6	10	1	8	9	5	2

Table 7. *Parameter estimates for corn data*

σ_u^2	ρ	α	β	σ^2
0	1.000	73.15	0.344	46.9
10	0.964	72.24	0.357	45.7
20	0.928	71.26	0.371	44.4
30	0.892	70.20	0.386	42.9
40	0.856	69.05	0.402	41.4
50	0.820	67.80	0.420	39.7
57	0.794	66.86	0.433	38.4
60	0.783	66.44	0.439	37.8
70	0.747	64.94	0.460	35.8
80	0.711	63.29	0.484	33.6
90	0.675	61.46	0.510	31.1
100	0.639	59.42	0.539	28.4

We give the ML analysis in Table 7 for a normal true-score distribution. The parameter estimates for $\sigma_u^2 = 57$ are close to those of Fuller ($\alpha = 67.56$, $\beta = 0.423$, $\sigma^2 = 43.3$); the difference in the variances is the difference between ML and REML approaches.

Estimating the true-score distribution nonparametrically, we find a four-point distribution for the true-score, and a very small error variance estimate: the parameter estimates are essentially identical to the least squares estimates. This is surprising as the measurement error variance used by Fuller is $\sigma_u^2 = 57$, i.e. $\sigma_u = 7.55$. An (X, Y) plot of the data, or simple inspection of the data table above, immediately shows why: the observations fall into four tightly clustered groups on X; the marginal distribution of X is certainly not normal.

Examination of the weights w_{ik} provides the same information: all the estimated weights are 1.000. The estimated true scores are 51.3, 64.0, 69.7 and 95.3, with the first three sites in the table assigned to the first true-score, the next two sites to the second, the next three to the third, and the last three to the fourth. Thus if we have no external information about the measurement error, its estimate is essentially the variation among the "near-replicate" values of nitrogen.

It is clear that external information about the error variance plays a critical role in parameter estimation, and a sensitivity analysis as above is an important adjunct to any analysis based on an assumed *or* an estimated error variance.

11. True-score estimation for the generalized linear model

Estimation of the true scores in the GLM is somewhat more complicated. Consider the general model with log density

$$\log f(y_i \mid \beta, \phi) = (y_i\theta_i - b(\theta_i))/\phi + c(y_i, \phi)$$

with mean μ and link function $g : \eta_i = g(\mu_i) = \alpha + \beta X_i + \gamma' z_i$, where X is the variable measured with error as W. As shown in §3, ML estimation of the parameters is a weighted form of the usual GLM analysis, and we do not discuss it further. Estimation of the π_k is as in §3 a standard mixture ML result.

For the true scores, the first derivative of the log-likelihood with respect to X_k can be expressed as

$$\begin{aligned}\frac{\partial \ell}{\partial X_k} &= \sum_i w_{ik}\left\{\frac{\partial \log f_{ik}}{\partial X_k} + \frac{\partial \log m_{ik}}{\partial X_k}\right\} \\ &= \sum_i w_{ik}\left\{a'(\eta_{ik})\frac{\beta}{\phi}[y_i - b'(\theta_{ik})]\right. \\ &\qquad \left. + \frac{1}{\sigma_u^2}[w_i - \mu - \sigma_x X_k - \lambda' z_i]\right\}\end{aligned}$$

where $a(\eta) = \theta$ is equal to the identity function if g is the canonical link. The Hessian of ℓ with respect to the X_k is a diagonal matrix having elements

$$\begin{aligned}\frac{\partial^2 \ell}{\partial X_k^2} = \sum_i w_{ik}\left\{\frac{\beta^2}{\phi}[a''(\eta_{ik})[y_i - b'(\theta_{ik})]\right. \\ \left. - \frac{\beta^2}{\phi}[a'(\eta_{ik})]^2 b''(\theta_{ik}) - \frac{\sigma_x}{\sigma_u^2}\right\}.\end{aligned}$$

In a Newton update, the new values for the X_k can be computed as

$$X_k^{(new)} = X_k^{(old)} - s\frac{\partial \ell}{\partial X_k} \Big/ \frac{\partial^2 \ell}{\partial X_k^2}$$

where the value of s (equal for each $k = 1, \ldots, K$) is found using a simple line search algorithm to increase the likelihood. For the special case of the natural link the two derivatives simplify to

$$\begin{aligned}\frac{\partial \ell}{\partial X_k} &= \sum_i w_{ik}\left\{\frac{\beta_1}{\phi}[y_i - b'(\theta_{ik})] + \frac{1}{\sigma_u^2}(w_i - X_k - \gamma' z_i)\right\} \\ \frac{\partial^2 \ell}{\partial X_k^2} &= \sum_i w_{ik}\left\{-\frac{\beta_1^2}{\phi}b''(\theta_{ik}) - \frac{1}{\sigma_u^2}\right\}\end{aligned}$$

A full discussion of the general model will appear elsewhere.

We finally extend the results above to the general covariate case.

12. Multiple error-prone explanatory variables

The previous analysis can be extended straightforwardly to multiple variables measured with error. We illustrate with the normal model with two error-prone variables, W_1 and W_2, with true scores X_1 and X_2. Further extensions will be obvious. The true scores have an unknown joint distribution $\pi(X_1, X_2)$. This is estimated by NPML as a discrete distribution on a finite number K of masspoints (X_{1k}, X_{2k}) with masses π_k. We assume that the measurement errors in these variables are independent (though correlated measurement errors can be allowed in a further extension) and normal, and the remaining explanatory variables Z are measured without error. As in §2 the reparametrised model

is then

$$Y \mid X_1, X_2, Z \sim f(Y \mid \eta, \phi)$$
$$W_1 \mid X_1, Z \sim N\big(X_1 + \lambda_1' Z, \sigma_{u1}^2\big)$$
$$W_2 \mid X_2, Z \sim N\big(X_2 + \lambda_2' Z, \sigma_{u2}^2\big)$$

with $\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3' Z$, and Y is conditionally independent of W_1 and W_2 given X_1, X_2 and Z.

The mixture likelihood to be maximized is now

$$L = \prod_{i=1}^{n} \left(\sum_{k=1}^{K} \pi_k f(y_i \mid X_{1k}, X_{2k}, z_i) \times m_1(w_{1i} \mid X_{1k}, z_i) m_2(w_{2i} \mid X_{2k}, z_i) \right)$$
$$= \prod_i \left(\sum_k \pi_k f_{ik} m_{1ik} m_{2ik} \right)$$

where m_1 and m_2 are the measurement densities.

Maximization of L with respect to β, λ_1, λ_2, ϕ, σ_{u1}^2, σ_{u2}^2 and π_k proceeds as in §§7 and 8. For the true scores X_{1k} and X_{2k}, we have on differentiation

$$\frac{\partial \ell}{\partial X_{1k}} = \sum_i w_{ik} \left[\frac{\beta_1}{\sigma^2}(y_i - \beta_1 X_{1k} - \beta_2 X_{2k} - \beta_3' z_i) + \frac{1}{\sigma_{u1}^2}(w_{1i} - X_{1k} - \lambda_1' z_i) \right]$$
$$\frac{\partial \ell}{\partial X_{2k}} = \sum_i w_{ik} \left[\frac{\beta_2}{\sigma^2}(y_i - \beta_1 X_{1k} - \beta_2 X_{2k} - \beta_3' z_i) + \frac{1}{\sigma_{u2}^2}(w_{2i} - X_{2k} - \lambda_2' z_i) \right]$$

where the weights w_{ik} are now given by

$$w_{ik} = \frac{\pi_k f_{ik} m_{1ik} m_{2ik}}{\sum_\ell \pi_\ell f_{i\ell} m_{1i\ell} m_{2i\ell}}.$$

This gives two linear equations in X_{1k} and X_{2k} which are easily solved:

$$X_{1k}\left(\frac{\beta_1^2}{\sigma^2} + \frac{1}{\sigma_{1u}^2}\right) + X_{2k}\left(\frac{\beta_1\beta_2}{\sigma^2}\right) = \frac{\beta_1}{\sigma^2}(\bar{y}_k - \beta_3' \bar{z}_k) + \frac{1}{\sigma_{1u}^2}(\bar{w}_{1k} - \lambda_1' \bar{z}_k)$$
$$X_{1k}\left(\frac{\beta_1\beta_2}{\sigma^2}\right) + X_{2k}\left(\frac{\beta_2^2}{\sigma^2} + \frac{1}{\sigma_{2u}^2}\right) = \frac{\beta_2}{\sigma^2}(\bar{y}_k - \beta_3' \bar{z}_k) + \frac{1}{\sigma_{2u}^2}(\bar{w}_{2k} - \lambda_2' \bar{z}_k)$$

where

$$\bar{y}_k = \sum_i w_{ik} y_i, \quad \bar{w}_{jk} = \sum_i w_{ik} w_{ji}, \quad j = 1, 2.$$

A full discussion of the general model will be presented elsewhere.

13. Discussion

The method presented here is very general. Particular strengths of the approach are that error-free covariates can be incorporated, and that no external or internal measure of reliability is required for the error-prone covariate(s) except for the case of normal true and error scores. Known measurement error variances can be incorporated simply, and substantially increase the amount of information in the data about the regression model parameters; they may have a considerable effect on parameter estimates. Known reliabilities are more difficult to incorporate and are not discussed here.

The discrete nature of the true-score distribution may appear strange, and many find a continuous true score distribution more appealing. Continuous true-score models (and continuous random effect models generally) can be fitted with minimal distributional assumptions by representing the true-score distribution as an unknown mixture of normals with the same standard deviation, giving a form of kernel density estimate of the mixing distribution (Magder and Zeger 1996), but little information is in general available about the shape of this distribution, since the likelihood is nearly flat in the "bandwidth" (common standard deviation) parameter of the kernel estimate (Aitkin 1999).

The approach followed here allows the identification of outliers or actual latent class structure in the data, which is not possible by continuous modelling. Aitkin (1996) gave an example of time sequencing identified in this way.

Provided that a discrete estimated true-score distribution is not "reified" without external supporting evidence, and is regarded simply as the best nonparametric guess at the true distribution, the discrete estimate is unlikely to be misleading. The NPML approach is minimalist, yet without assumptions we still obtain full ML estimates.

The EM algorithm implementation used here is relatively straightforward to program, though slow; faster gradient algorithms for mixture models were described by Lesperance and Kalbfleisch (1992).

References

Aitkin M. 1996. A general maximum likelihood analysis of overdispersion in generalized linear models. Statistics and Computing 6: 251–262.

Aitkin M., Anderson D.A., Francis B.J., and Hinde J.P. 1989. Statistical Modelling in GLIM. University Press, Oxford.

Aitkin M. and Francis B.J. 1995. Fitting overdispersed generalized linear models by nonparametric maximum likelihood. The GLIM Newsletter 25: 37–45.

Carroll R.J., Ruppert D., and Stefanski L.A. 1995. Measurement Error in Nonlinear Models. Chapman and Hall, London.

Carroll R.J. and Stefanski L.A. 1990. Approximate quasilikelihood estimation in models with surrogate predictors. J. Amer. Statist. Assoc. 85: 652–663.

Davies R.B. 1987. Mass point methods for dealing with nuisance parameters in longitudinal studies. In: R. Crouchley, (Ed.), Longitudinal Data Analysis. Avebury, Aldershot, Hants.

Dempster A.P., Laird N.M., and Rubin D.A. 1977. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). J. Roy. Statist. Soc. B 39: 1–38.

Fuller W. 1987. Measurement Error Models. John Wiley, New York.

Gleser L.J. 1990. Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models. In: P.J. Brown and W.A. Fuller (Eds.), Statistical Anaysis of Measurement Error Models and Application, American Mathematics Society, Providence.

Hanfelt J.J. and Liang K.-Y. 1997. Approximate likelihoods for generalized errors-in-variables models. J. Roy. Statist. Soc. B 59: 627–637.

Heckman J.J and Singer B. 1984. A method for minimizing the impact of distributional assumptions in econometric models of duration. Econometrica 52: 271–320.

Hinde J.P. 1982. Compound Poisson regression models. In: R. Gilchrist, (Ed.), GLIM 82 , Springer-Verlag, New York.

Hinde J.P. and Wood A.T.A. 1987. Binomial variance component models with a non-parametric assumption concerning random effects. In: R. Crouchley (Ed.), Longitudinal Data Analysis. Avebury, Aldershot, Hants.

Kiefer J. and Wolfowitz J. 1956. Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters. Ann. Math. Statist. 27: 887–906.

Laird N.M. 1978. Nonparametric maximum likelihood estimation of a mixing distribution. J. Amer. Statist. Assoc. 73: 805–811.

Lesperance M.L. and Kalbfleisch J.D. 1992. An algorithm for computing the nonparametric MLE of a mixing distribution. J. Amer. Statist. Assoc. 87: 120–126.

Lindsay B.G. 1983. The geometry of mixture likelihoods, part I: A general theory. Ann. Statist. 11: 86–94.

McLachlan G.J. 1987. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. Applied Statistics 36: 318–324.

Magder L.S. and Zeger S.L. 1996. A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. J. Amer. Statist. Assoc. 91: 1141–1151.

Murphy S.A. and Van Der Vaart A.W. 2000. On profile likelihoods (with discussion). J. Amer. Statist. Assoc. 95: 449–485.

Nakamura T. 1990. Corrected score functions for errors-in-variables models: Methodology and application to generalized linear models. Biometrika 77: 127–137.

Reiersøl O. 1950. Identifiability of a linear relation between variables which are subject to error. Econometrica 18: 375–389.

Roeder K., Carroll R.J., and Lindsay B.G. 1996. A semiparametric mixture approach to case-control studies with errors in covariables. J. Amer. Statist. Assoc. 91: 722–732.

Schafer D.W. 1987. Covariate measurement error in generalized linear models. Biometrika 74: 385–391.