Statistica Neerlandica (2004) Vol. 58, nr. 2, pp. 220–233

An EM algorithm for the estimation of parametric and nonparametric hierarchical nonlinear models

Jeroen K. Vermunt*

Department of Methodology and Statistics, Tilburg University,
PO Box 90153, 5000 LE Tilburg

It is shown how to implement an EM algorithm for maximum likelihood estimation of hierarchical nonlinear models for data sets consisting of more than two levels of nesting. This upward–downward algorithm makes use of the conditional independence assumptions implied by the hierarchical model. It cannot only be used for the estimation of models with a parametric specification of the random effects, but also to extend the two-level nonparametric approach – sometimes referred to as latent class regression – to three or more levels. The proposed approach is illustrated with an empirical application.

Key Words and Phrases: nonlinear random-effects model, nonlinear mixed model, multilevel analysis, latent class analysis, finite mixture model, latent class regression, maximum likelihood estimation.

1 Introduction

A well-established estimation method for hierarchical models is maximum likelihood (ML). While ML estimation is straightforward with normal level-1 errors, with nonnormal dependent variables, it requires approximation of the integrals in the likelihood function corresponding to the mixing distribution. The most common method is to approximate the likelihood function using Gauss–Hermite or adaptive quadrature numerical integration. Software packages implementing such methods include the MIXOR family programs (HEDEKER and GIBBONS, 1996), the SAS NLMIXED procedure, and the STATA GLLAMM routine (RABE-HESKETH, PICKLES and SKRONDAL 2001, 2002, 2003). MIXOR and NLMIXED are two-level programs. GLLAMM can also be used for ML estimation of hierarchical nonlinear models with more than two levels of nesting.

If the mixed model of interest is a two-level model, ML estimation can be performed by means of the EM algorithm (BOCK and AITKIN, 1981; AGRESTI *et al.*, 2000), which is a natural approach to estimation problems with missing data (here, the random effects

*j.k.vermunt@uvt.nl

I would like to thank the two reviewers and Sophia Rabe-Hesketh for their comments on an earlier version of the paper.

for each case). The standard EM algorithm can, however, not be used for other types of mixed models because the number of entries in the relevant posterior distribution is huge, making the method impractical. To be more specific about the problem associated with standard EM, suppose that we have an three-level random-intercept model. The E step involves computing the posterior distribution of the random intercepts corresponding to each of the three-levels units; that is, the joint distribution of the level-3 random intercept and the random intercepts for all level-2 units belonging to the level-3 unit concerned. With 10 quadrature nodes and 20 level-2 units per level-3 unit, this is a posterior distribution with 10^{1+20} entries. This illustrates that computer storage and time increases exponentially with the number of level-2 units within level-3 units, which makes EM impractical with more than a few level-2 units per level-3 unit. This is unfortunate because EM is a very stable and quite fast algorithm.

MIXOR, NLMIXED, and GLLAMM maximize the log-likelihood using Newton-type algorithms. Lesaffre and Spiessens (2001) reported difficulties with the MIXOR and NLMIXED Newton-type algorithms in finding the global ML solution in nonlinear mixed (two-level) models: different routines and algorithms may give different solutions for the same number of quadrature points. It can be expected that this problem becomes even worse in models with more than two levels. Fortunately, the Newton algorithm used by GLLAMM – the only one of the three programs that can deal with more than two levels of nesting – seems to be more stable. GLLAMM, however, uses numerical first and second derivatives of the log-likelihood, which is computationally intensive in models with more than a few parameters. Although it cannot be expected that EM resolves all the "problems" associated with the Newton-type methods, it would be useful to have an EM algorithm for nonlinear hierarchical models as an additional tool. The most important advantage of EM is that it converges irrespective of the starting values. Analytical derivatives for the M step are readily obtained, but can also be adopted from existing generalized linear modeling packages.

The most common specification for the mixing distribution is multivariate normal. However, instead of working with such a parametric distribution for the random coefficients, it is also possible to use a nonparametric specification (Laird, 1978; Aitkin, 1999). This yields what is usually referred to as latent class regression or finite mixture regression (Vermunt and Magidson, 2000; Vermunt and Van Dijk, 2001; Wedel and DeSarbo, 2002). An advantage of such a nonparametric approach is that it is not necessary to introduce possibly inappropriate and unverifiable assumptions about the distribution of the random effects (Aitkin, 1999).

The latent class regression models developed so far can, however, only deal with two-level data structures. The main reason for this is that, as explained above for the parametric case, the standard EM algorithm cannot be used for ML estimation of nonparametric hierarchical models with more than two levels. Although it is possible to estimate latent classes regression models using Newton–Raphson, it is well-known that this requires good starting values and that such good starting values may be difficult to find.

This paper shows how to implement an EM algorithm for ML estimation of parametric and nonparametric hierarchical nonlinear models with more than two levels. The new algorithm makes use of the conditional independence assumptions implied by the hierarchical model of interest. More specifically, it is based on the fact that lower level observations are independent of each other given the higher-level random effects. The underlying idea of using the structure of the model of interest for the implementation of the EM algorithm is similar to what is done in hidden Markov models. For these models, BAUM *et al.* (1970) developed an efficient EM algorithm which is known as the forward–backward algorithm because it moves forward and backward through the hidden Markov chain. The version of EM described in this paper will be called the upward–downward algorithm because it moves upward and downward through the hierarchical structure. VERMUNT (2003) used the same kind of algorithm for a two-level extension of the standard latent model. Although for simplicity of exposition I will concentrate on the three-level case, the proposed method can easily be generalized to any number of levels.

The next section describes the parametric three-level hierarchical model of interest, as well as its ML estimation by means of the upward–downward algorithm. Subsequently, attention is paid to the three-level extension of the latent class regression model. The proposed methods are illustrated with an empirical application. The paper ends with a short discussion.

2 The nonlinear three-level model with parametric random effects

Let i denote a level-1 unit, j a level-2 unit, and k a level-3 unit. The total number of level-3 units is denoted by K, the number of level-2 units within level-3 unit k by n_k, and the number of level-1 units within level-2 unit jk by n_{jk}. Let y_{ijk} be the response of level-1 unit ijk on the outcome variable of interest, and let $\mathbf{x}_{ijk}$, $\mathbf{z}_{ijk}^{(2)}$, and $\mathbf{z}_{ijk}^{(3)}$ be the design vectors associated with S fixed effects, $R^{(2)}$ level-2 random effects, and $R^{(3)}$ level-3 random effects, respectively. It is assumed that the conditional densities of the responses given covariates and random effects are from the exponential family. Denoting the link function by $g[\cdot]$, the nonlinear three-level model (NLTM) can be defined as

$$g[E(y_{ijk}|\mathbf{x}_{ijk}, \mathbf{z}_{ijk}^{(2)}, \mathbf{z}_{ijk}^{(3)}, \boldsymbol{\beta}_{jk}^{(2)}, \boldsymbol{\beta}_{k}^{(3)})] = \eta_{ijk} = \mathbf{x}'_{ijk}\boldsymbol{\alpha} + \mathbf{z}_{ijk}^{(2)\prime}\boldsymbol{\beta}_{jk}^{(2)} + \mathbf{z}_{ijk}^{(3)\prime}\boldsymbol{\beta}_{k}^{(3)}.$$

Here, $\boldsymbol{\alpha}$ is the vector of unknown fixed effects, $\boldsymbol{\beta}_{jk}^{(2)}$ is the vector of unknown random effects for level-2 unit jk, and $\boldsymbol{\beta}_{k}^{(3)}$ is the vector of unknown random effects for level-3 unit k.

As usual, we assume the distribution of the random effects $\boldsymbol{\beta}_{jk}^{(2)}$ and $\boldsymbol{\beta}_{k}^{(3)}$ to be multivariate normal with zero mean vector and covariance matrices $\boldsymbol{\Sigma}^{(2)}$ and $\boldsymbol{\Sigma}^{(3)}$. For parameter estimation, it is convenient to standardize and orthogonalize the random effects. For this, let $\boldsymbol{\beta}_{jk}^{(2)} = \mathbf{C}^{(2)}\boldsymbol{\theta}_{jk}^{(2)}$, where $\mathbf{C}^{(2)}\mathbf{C}^{(2)\prime}=\boldsymbol{\Sigma}^{(2)}$ is the Cholesky decomposition of $\boldsymbol{\Sigma}^{(2)}$. Similarly, we define $\boldsymbol{\beta}_{k}^{(3)} = \mathbf{C}^{(3)}\boldsymbol{\theta}_{k}^{(3)}$. The reparameterized NLTM is then

$$\eta_{ijk} = \mathbf{x}'_{ijk}\boldsymbol{\alpha} + \mathbf{z}^{(2)\prime}_{ijk}\mathbf{C}^{(2)}\boldsymbol{\theta}^{(2)}_{jk} + \mathbf{z}^{(3)\prime}_{ijk}\mathbf{C}^{(3)}\boldsymbol{\theta}^{(3)}_{k}. \tag{1}$$

The means and variances of $\boldsymbol{\theta}^{(2)}_{jk}$ and $\boldsymbol{\theta}^{(3)}_{k}$ are 0 and 1, respectively. Note that $\boldsymbol{\alpha}$, $\mathbf{C}^{(2)}$, and $\mathbf{C}^{(3)}$ contain the unknown parameters to be estimated.

Log-likelihood function

The parameters of the NLTM can be estimated by maximum likelihood (ML). The likelihood function is based on the probability densities of the level-3 observations, denoted by $P(\mathbf{y}_k|\mathbf{x}_k, \mathbf{z}^{(2)}_k, \mathbf{z}^{(3)}_k)$. Here, $\mathbf{y}_k$, $\mathbf{x}_k$, $\mathbf{z}^{(2)}_k$, and $\mathbf{z}^{(3)}_k$ contain the responses and design vectors for all cases belonging to level-3 unit k. In order to simplify notation, the conditioning on the design vectors is replaced by an index corresponding to the unit concerned, yielding the short-hand notation $P_k(\mathbf{y}_k)$ for the probability density of level-3 unit k.

The log-likelihood to be maximized equals

$$\log L = \sum_{k=1}^{K} \log P_k(\mathbf{y}_k),$$

where

$$\begin{aligned} P_k(\mathbf{y}_k) &= \int_{\boldsymbol{\theta}^{(3)}} P_k(\mathbf{y}_k|\boldsymbol{\theta}^{(3)}) f(\boldsymbol{\theta}^{(3)}) \mathrm{d}\boldsymbol{\theta}^{(3)} \\ &= \int_{\boldsymbol{\theta}^{(3)}} \left\{ \prod_{j=1}^{n_k} P_{jk}(\mathbf{y}_{jk}|\boldsymbol{\theta}^{(3)}) \right\} f(\boldsymbol{\theta}^{(3)}) \mathrm{d}\boldsymbol{\theta}^{(3)}, \end{aligned} \tag{2}$$

and

$$\begin{aligned} P_{jk}(\mathbf{y}_{jk}|\boldsymbol{\theta}^{(3)}) &= \int_{\boldsymbol{\theta}^{(2)}} P_{jk}(\mathbf{y}_{jk}|\boldsymbol{\theta}^{(2)}, \boldsymbol{\theta}^{(3)}) f(\boldsymbol{\theta}^{(2)}) \mathrm{d}\boldsymbol{\theta}^{(2)} \\ &= \int_{\boldsymbol{\theta}^{(2)}} \left\{ \prod_{i=1}^{n_{jk}} P_{ijk}(y_{ijk}|\boldsymbol{\theta}^{(2)}, \boldsymbol{\theta}^{(3)}) \right\} f(\boldsymbol{\theta}^{(2)}) \mathrm{d}\boldsymbol{\theta}^{(2)}. \end{aligned} \tag{3}$$

As can be seen, the responses of the n_k level-2 units within level-3 unit k are assumed to be independent of one another given the random effects $\boldsymbol{\theta}^{(3)}$, and the responses of the n_{jk} level-1 units within level-2 unit jk are assumed to be independent of one another given the random effects $\boldsymbol{\theta}^{(2)}$ and $\boldsymbol{\theta}^{(3)}$. Note that level-2 and level-3 random effects are assumed to be mutually independent – $f(\boldsymbol{\theta}^{(2)}|\boldsymbol{\theta}^{(3)}) = f(\boldsymbol{\theta}^{(2)})$ – which is a common assumption in multilevel models.

The integrals at the right-hand side of equations (2) and (3) can be evaluated by the Gauss–Hermite quadrature numerical integration method (Stroud and Secrest, 1966; Bock and Aitkin, 1981; Hedeker and Gibbons, 1996; Rabe-Hesketh, Pickles and Skrondal 2001, 2002), in which the multivariate normal mixing distribution is approximated by a limited number of discrete points. More precisely, the integrals are replaced by summations over M and T quadrature points,

$$\begin{aligned}P_k(\mathbf{y}_k) &= \sum_{m=1}^{M} P_k(\mathbf{y}_k|\boldsymbol{\theta}_m^{(3)})\pi(\boldsymbol{\theta}_m^{(3)})\\ &= \sum_{m=1}^{M}\left[\prod_{j=1}^{n_k} P_{jk}(\mathbf{y}_{jk}|\boldsymbol{\theta}_m^{(3)})\right]\pi(\boldsymbol{\theta}_m^{(3)})\\ &= \sum_{m=1}^{M}\left[\prod_{j=1}^{n_k}\sum_{t=1}^{T}\left\{\prod_{i=1}^{n_{jk}} P_{ijk}(y_{ijk}|\boldsymbol{\theta}_t^{(2)},\boldsymbol{\theta}_m^{(3)})\right\}\pi(\boldsymbol{\theta}_t^{(2)})\right]\pi(\boldsymbol{\theta}_m^{(3)}).\end{aligned} \tag{4}$$

Actually, we should use a "$\approx$" instead of a "$=$" sign in this expression because we are approximating the integral by a summation. However, for simplicity of notation in this and next formulas, we retain the "$=$".

In the above formula, $\boldsymbol{\theta}_t^{(2)}$ and $\boldsymbol{\theta}_m^{(3)}$ are quadrature nodes and $\pi(\boldsymbol{\theta}_t^{(2)})$ and $\pi(\boldsymbol{\theta}_m^{(3)})$ are quadrature weights corresponding to the (multivariate) normal densities of interest. Because the random effects are orthogonalized, the nodes and weights of the separate dimensions equal the ones of the univariate normal density, which can be obtained from standard tables (see, for example, STROUD and SECREST, 1966). Suppose that each dimension is approximated with Q quadrature nodes. The $T = Q^{R^{(2)}}$ and $M = Q^{R^{(3)}}$ weights are then obtained by multiplying the weights of the separate dimensions. The integral can be approximated to any practical degree of accuracy by setting Q sufficiently large. LESAFFRE and SPIESSENS (2001) and RABE-HESKETH, SKRONDAL and PICKLES (2002) showed that the number of quadrature points needs to be very large in some situations. In such cases, it is better to use adaptive quadrature.

The upward–downward variant of the EM algorithm

A natural way to solve the ML estimation problem of the parameters $\boldsymbol{\alpha}$, $\mathbf{C}^{(2)}$, and $\mathbf{C}^{(3)}$ is by means of the EM algorithm (DEMPSTER, LAIRD and RUBIN, 1977). The E step of the EM algorithm involves computing the expectation of the complete data log-likelihood, which in the NLTM is of the form

$$\log L_c = \sum_{m=1}^{M}\sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}} P_{jk}(\boldsymbol{\theta}_t^{(2)},\boldsymbol{\theta}_m^{(3)}|\mathbf{y}_k)\log P_{ijk}(y_{ijk}|,\boldsymbol{\theta}_t^{(2)},\boldsymbol{\theta}_m^{(3)}). \tag{5}$$

The terms containing the priors $\pi(\boldsymbol{\theta}_t^{(2)})$ and $\pi(\boldsymbol{\theta}_m^{(3)})$ are omitted from L_c because these do not contain parameters to be estimated.

Equation (5) shows that, in fact, the E step involves obtaining the posterior probabilities $P_{jk}(\boldsymbol{\theta}_t^{(2)},\boldsymbol{\theta}_m^{(3)}|\mathbf{y}_k)$ given the current estimates for the unknown model parameters. In the M step of the algorithm, the $\boldsymbol{\alpha}$, $\mathbf{C}^{(2)}$, and $\mathbf{C}^{(3)}$ parameters are updated so that the expected complete data log-likelihood given in equation (5) is maximized (or improved). This can be accomplished using standard algorithms for the ML estimation of generalized linear models.

The problematic part in the implementation of EM for the NLTM is the E step in which one has to obtained the posterior probabilities $P_{jk}(\boldsymbol{\theta}_t^{(2)},\boldsymbol{\theta}_m^{(3)}|\mathbf{y}_k)$. A standard implementation of the E step would involve computing the joint conditional expectation of the $n_k{\cdot}R^{(2)}+R^{(3)}$ random effects for level-3 unit k; that is, the joint posterior distribution $P_k(\boldsymbol{\theta}_{t_1}^{(2)}, \boldsymbol{\theta}_{t_2}^{(2)}, \ldots, \boldsymbol{\theta}_{t_{n_k}}^{(2)}, \boldsymbol{\theta}_m^{(3)}|\mathbf{y}_k)$ with $M{\cdot}T^{n_k}$ entries. Note that this amounts to computing the expectation of all the "missing data" for a level-3 unit. These joint posteriors would subsequently be collapsed to obtain the marginal posterior probabilities for each level-2 unit j within level-3 unit k, $P_{jk}(\boldsymbol{\theta}_t^{(2)}, \boldsymbol{\theta}_m^{(3)}|\mathbf{y}_k)$. This yields a procedure in which computer storage and time increases exponentially with the number of level-2 units, which means that it can only be used with very small n_k.

However, it turns out that it is possible to compute the n_k marginal posterior probability distributions $P_{jk}(\boldsymbol{\theta}_t^{(2)}, \boldsymbol{\theta}_m^{(3)}|\mathbf{y}_k)$ without going through the full posterior distribution by making use of the conditional independence assumptions associated with the density function defined in equation (2). In that sense, our procedure is similar to the forward–backward algorithm for the estimation of hidden Markov models with large numbers of time points (BAUM *et al.*, 1970; JUANG and RABINER, 1991). The procedure described below could be called an upward–downward algorithm. First, random effects are integrated out, going from the lower to the higher levels. Subsequently, the relevant marginal posterior probabilities are computed going from the higher to the lower levels. This yields a procedure in which computer storage and time increases only linearly with the number of level-2 observations instead of exponentially, as would have been the case with a standard EM algorithm.

The marginal posterior probabilities $P_{jk}(\boldsymbol{\theta}_t^{(2)}, \boldsymbol{\theta}_m^{(3)}|\mathbf{y}_k)$ can be decomposed as follows:

$$P_{jk}(\boldsymbol{\theta}_t^{(2)}, \boldsymbol{\theta}_m^{(3)}|\mathbf{y}_k) = P_k(\boldsymbol{\theta}_m^{(3)}|\mathbf{y}_k)P_{jk}(\boldsymbol{\theta}_t^{(2)}|\mathbf{y}_k, \boldsymbol{\theta}_m^{(3)}).$$

Our procedure makes use of the fact that in the NLTM

$$P_{jk}(\boldsymbol{\theta}_t^{(2)}|\mathbf{y}_k, \boldsymbol{\theta}_m^{(3)}) = P_{jk}(\boldsymbol{\theta}_t^{(2)}|\mathbf{y}_{jk}, \boldsymbol{\theta}_m^{(3)});$$

i.e., $\boldsymbol{\theta}_t^{(2)}$ is independent of the observed and latent variables of the other level-2 units within the same level-3 unit given $\boldsymbol{\theta}^{(3)}$. This is the result of the fact that level-2 observations are mutually independent given the level-3 random effects, as is expressed in the density function described in equation (2). Using this important result, we get the following slightly simplified decomposition:

$$P_{jk}(\boldsymbol{\theta}_t^{(2)}, \boldsymbol{\theta}_m^{(3)}|\mathbf{y}_k) = P_k(\boldsymbol{\theta}_m^{(3)}|\mathbf{y}_k)P_{jk}(\boldsymbol{\theta}_t^{(2)}|\mathbf{y}_{jk}, \boldsymbol{\theta}_m^{(3)}). \tag{6}$$

The computation of the marginal posterior probabilities therefore reduces to the computation of the two terms on the right-hand side of this equation. The term $P_k(\boldsymbol{\theta}_m^{(3)}|\mathbf{y}_k)$ is obtained by

$$P_k(\boldsymbol{\theta}_m^{(3)}|\mathbf{y}_k) = \frac{P_k(\mathbf{y}_k, \boldsymbol{\theta}_m^{(3)})}{P_k(\mathbf{y}_k)} \tag{7}$$

where

$$P_k(\mathbf{y}_k, \boldsymbol{\theta}_m^{(3)}) = \pi(\boldsymbol{\theta}_m^{(3)}) \prod_{j=1}^{n_k} P_{jk}(\mathbf{y}_{jk}|\boldsymbol{\theta}_m^{(3)})$$

$$P_k(\mathbf{y}_k) = \sum_{m=1}^{M} P(\mathbf{y}_k, \boldsymbol{\theta}_m^{(3)}).$$

The other term, $P_{jk}(\boldsymbol{\theta}_t^{(2)}|\mathbf{y}_{jk}, \boldsymbol{\theta}_m^{(3)})$, is computed by

$$P_{jk}(\boldsymbol{\theta}_t^{(2)}|\mathbf{y}_{jk}, \boldsymbol{\theta}_m^{(3)}) = \frac{P_{jk}(\mathbf{y}_{jk}, \boldsymbol{\theta}_t^{(2)}|\boldsymbol{\theta}_m^{(3)})}{P_{jk}(\mathbf{y}_{jk}|\boldsymbol{\theta}_m^{(3)})},$$

where

$$P_{jk}(\mathbf{y}_{jk}, \boldsymbol{\theta}_t^{(2)}|\boldsymbol{\theta}_m^{(3)}) = \pi(\boldsymbol{\theta}_t^{(2)}) \prod_{i=1}^{n_{jk}} P_{ijk}(y_{ijk}|\boldsymbol{\theta}_t^{(2)}, \boldsymbol{\theta}_m^{(3)})$$

$$P_{jk}(\mathbf{y}_{jk}|, \boldsymbol{\theta}_m^{(3)}) = \sum_{t=1}^{T} P_{jk}(\mathbf{y}_{jk}, \boldsymbol{\theta}_t^{(2)}|\boldsymbol{\theta}_m^{(3)}).$$

Thus, first the level-2 posterior probabilities $P_{jk}(\boldsymbol{\theta}_t^{(2)}|\mathbf{y}_{jk}, \boldsymbol{\theta}_m^{(3)})$ are obtained from the level-1 information $P_{ijk}(y_{ijk}|\boldsymbol{\theta}_t^{(2)}, \boldsymbol{\theta}_m^{(3)})$, and subsequently the level-3 posterior probabilities $P_k(\boldsymbol{\theta}_m^{(3)}|\mathbf{y}_k)$ are obtained from the level-2 information $P_{jk}(\mathbf{y}_{jk}|\boldsymbol{\theta}_m^{(3)})$. This is called the *upward* step of the algorithm because one goes up in the hierarchical structure. In the *downward* step, one computes $P_{jk}(\boldsymbol{\theta}_t^{(2)}, \boldsymbol{\theta}_m^{(3)}|\mathbf{y}_k)$ by means of equation (6).

The upward–downward method can easily be generalized to more than three levels. For example, with four levels, one would have to compute the three terms $P_\ell(\boldsymbol{\theta}_o^{(4)}|\mathbf{y}_\ell)$, $P_{k\ell}(\boldsymbol{\theta}_m^{(3)}|\mathbf{y}_{k\ell}, \boldsymbol{\theta}_o^{(4)})$, and $P_{jk\ell}(\boldsymbol{\theta}_t^{(2)}|\mathbf{y}_{jk\ell}, \boldsymbol{\theta}_m^{(3)}, \boldsymbol{\theta}_o^{(4)})$, where ℓ refers to a level-four unit and o to a quadrature point for the level-four unit random effects. These three terms are obtained in the upward step and used to calculate the relevant marginal posteriors in the downward step.

A practical problem in the implementation of the E step is that underflows may occur in the computation of $P_k(\boldsymbol{\theta}_m^{(3)}|\mathbf{y}_k)$. More precisely, the numerator of equation (7) may become equal to zero for each m because it may involve multiplication of a large number, $(n_k + 1)(n_{jk} + 1)$, of probabilities. Such underflows can, however, be prevented by working on a log scale. Letting $a_{mk} = \log[\pi(\boldsymbol{\theta}_m^{(3)})] + \sum_j^{n_k} \log[P_{jk}(\mathbf{y}_{jk}|\boldsymbol{\theta}_m^{(3)})]$ and $b_k = \max(a_{mk})$, $P_k(\boldsymbol{\theta}_m^{(3)}|\mathbf{y}_k)$ can be obtained by

$$P_k(\boldsymbol{\theta}_m^{(3)}|\mathbf{y}_k) = \frac{\exp(a_{mk} - b_k)}{\sum_p^M \exp(a_{pk} - b_k)}.$$

Standard errors and identification issues

Unlike Newton-like methods, the EM algorithm does not provide standard errors of the model parameters as a by-product. Estimated asymptotic standard errors can be obtained by computing the observed information matrix, the matrix of second-order derivatives of the log-likelihood function toward all model parameters. The inverse of this matrix is the estimated variance–covariance matrix. For the example presented later on, the necessary second derivatives were obtained numerically using analytic first derivatives. Note that the first derivatives are provided by the proposed EM algorithm.

The information matrix can also be used to check identifiability. A sufficient condition for local identification is that all the eigenvalues of this matrix are larger than zero. Although it is based on limited experience, so far no identification problems were encountered in the NLTMs that were estimated.

3 The nonlinear three-level model with nonparametric random effects

So far, we assumed that the random effects at the various levels come from parametric distributions. It is, however, also possible to work with discrete unspecified mixing distributions yielding a nonparametric random-effects approach (LAIRD, 1978). For the two-level case, these models are usually referred to as latent class or finite mixture regression models (VERMUNT and MAGIDSON, 2000; VERMUNT and VAN DIJK, 2001; WEDEL and DESARBO, 2002). Here, I present a three-level extension but, as in the parametric case, extension to more than three levels is straightforward.

An advantage of the presented nonparametric approach is that it is not necessary to introduce possibly inappropriate and unverifiable assumptions about the distribution of the random effects (AITKIN, 1999). Another important advantage is that it is computationally much less intensive than the parametric approach, especially in models containing more than two or three random effects.

Using the same notation as in the previous section, a three-level latent class regression model could be specified as follows

$$g[E(y_{ijk}|\mathbf{x}_{ijk}, \mathbf{z}^{(2)}_{ijk}, \mathbf{z}^{(3)}_{ijk}, \boldsymbol{\beta}^{(2)}_t, \boldsymbol{\beta}^{(3)}_m)] = \eta_{ijk|tm} = \mathbf{x}'_{ijk}\boldsymbol{\alpha} + \mathbf{z}^{(2)\prime}_{ijk}\boldsymbol{\beta}^{(2)}_t + \mathbf{z}^{(3)\prime}_{ijk}\boldsymbol{\beta}^{(3)}_m.$$

Here, $\boldsymbol{\alpha}$ is the vector of unknown fixed effects, $\boldsymbol{\beta}^{(2)}_t$ is the vector of unknown random effects for level-2 units belonging to latent class t, and $\boldsymbol{\beta}^{(3)}_m$ is the vector of unknown random effects for level-3 units belonging to latent class m. For identification, the parameters for $m = 1$ and $t = 1$ are fixed to zero, which amounts to using dummy coding for the "nominal" latent class variables.

As can be seen, an important difference with the parametric approach is that it is no longer assumed that each level-2 and each level-3 unit has its own set of regression parameters. Instead it is assumed that each level-2 unit belongs to one of T latent classes of level-2 units, and that each level-3 unit belongs to one of M latent classes of level-3 units. Each latent class has its own set of regression coefficients. With the maximum number of identifiable latent classes, the mixing distribution may be interpreted as a nonparametric distribution, yielding what is called the nonparametric ML estimator (NPMLE; LAIRD, 1978). In practice, however, we will stop increasing the number of latent classes when the model fit no longer improves.

The contribution to the likelihood function of the level-3 case k is similar to the contribution described in equation (4); that is,

$$P_k(\mathbf{y}_k) = \sum_{m=1}^{M} \left[\prod_{j=1}^{n_k} \sum_{t=1}^{T} \left\{ \prod_{i=1}^{n_{jk}} P_{ijk|tm}(y_{ijk}) \right\} \pi_t^{(2)} \right] \pi_m^{(3)}. \tag{8}$$

An important difference with the parametric case is that this is not an approximate density but an exact density. Moreover, the probabilities $\pi_t^{(2)}$ and $\pi_m^{(3)}$ are now unknown parameters to be estimated instead of fixed quadrature weights. The other unknown parameters determining the probabilities $P_{ijk|tm}(y_{ijk})$ are the fixed and class-specific regression coefficients $\boldsymbol{\alpha}$, $\boldsymbol{\beta}_t^{(2)}$, and $\boldsymbol{\beta}_m^{(3)}$.

The ML estimation problem of the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}_t^{(2)}$, $\boldsymbol{\beta}_m^{(3)}$, $\pi_t^{(2)}$ and $\pi_m^{(3)}$ can be solved by means of the EM algorithm (Dempster, Laird and Rubin, 1977). The E step of the EM algorithm involves computing the expectation of the complete data log-likelihood, which in the nonparametric NLTM is of the form

$$\begin{aligned} \log L_c = {} & \sum_{m=1}^{M} \sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{j=1}^{n_k} \sum_{i=1}^{n_{jk}} P_{jk}(t, m|\mathbf{y}_k) \log P_{ijk|tm}(y_{ijk}) \\ & + \sum_{m=1}^{M} \sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{j=1}^{n_k} P_{jk}(t, m|\mathbf{y}_k) \log \pi_t^{(2)} \\ & + \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{j=1}^{n_k} P_k(m|\mathbf{y}_k) \log \pi_m^{(3)}. \end{aligned} \tag{9}$$

This shows that, in fact, the E step involves obtaining the posterior probabilities $P_{jk}(t,m|\mathbf{y}_k)$ and $P_k(m|\mathbf{y}_k)$ given the current estimates for the unknown model parameters. In the M step of the algorithm, the model parameters are updated so that the expected complete data log-likelihood given in equation (9) is maximized (or improved). This can be accomplished using standard algorithms for the ML estimation of generalized linear models.

The upward–downward version of the EM algorithm proceeds in the same manner as in the parametric case. Instead of computing the $T \cdot M$ marginal posteriors $P_{jk}(\boldsymbol{\theta}_t^{(2)}, \boldsymbol{\theta}_m^{(3)}|\mathbf{y}_k)$ associated with the quadrature points, we have to obtain the $T \cdot M$ marginal posteriors $P_{jk}(t, m|\mathbf{y}_k)$; that is, the posterior probability that a level-2 unit j belongs to latent class t and level-3 unit k to latent class m. As can be seen from equation (7), the univariate posteriors $P_k(m|\mathbf{y}_k)$ are obtained as a by-product of the upward–downward algorithm.

4 Application to attitudes towards abortion data

To illustrate the NLTM, I obtained a data set from the data library of the Multilevel Models Project, at the Institute of Education, University of London (multilevel.ioe.ac.uk/intro/datasets.html). The data consist of 264 participants in 1983 to

1986 yearly waves from the British Social Attitudes Survey (McGrath and Waterton, 1986). It is a three-level data set: Individuals are nested within districts and time points are nested within individuals. The total number of level-3 units (districts) is 54.

The dependent variable is the number of yes responses on seven yes/no questions as to whether it is a woman's right to have an abortion under a specific circumstance. Because this variable is a count with a fixed total, it is most natural to work with a logit link and binomial error function. Individual level predictors in the data set are religion, political preference, gender, age, and self-assessed social class. In accordance with the results of Goldstein (1995), we found no significant effects of gender, age, self-assessed social class, and political preference. Therefore, we did not use these predictors in the further analysis. The predictors that were used are the level-1 predictor year of measurement (1 = 1983; 2 = 1984; 3 = 1985; 4 = 1986) and the level-2 predictor religion (1 = Roman Catholic, 2 = Protestant; 3 = Other; 4 = No religion). Because there was no evidence for a linear time effect, we included time as a set of dummies in the regression model.

The most general three-level model that is used contains a fixed intercept, 6 fixed slopes (three for time and three for religion), a random intercept at level 2, and a random intercept at level 3. The parametric form of this model is

$$\eta_{ijk} = \alpha_0 + \sum_{\ell=1}^{6} \alpha_\ell x_{\ell ijk} + \mathbf{c}^{(2)}\theta_j^{(2)} + \mathbf{c}^{(3)}\theta_k^{(3)},$$

where η_{ijk} is the logit of agreeing with an item. Note that $\mathbf{c}^{(2)}$ and $\mathbf{c}^{(3)}$ are the standard deviations of the two random intercepts. The nonparametric three-level model used is of the form

$$\eta_{ijk|tm} = \alpha_0 + \sum_{\ell=1}^{6} \alpha_\ell x_{\ell ijk} + \beta_t^{(2)} + \beta_m^{(3)}.$$

The analysis was performed with an experimental version of the Latent GOLD program (Vermunt and Magidson, 2000) that implements both the parametric and the nonparametric NLTM.

Table 1 reports fit measures obtained with the various models that were estimated. In the computation of BIC, I treated the number of level-3 units (54) as the total sample size. There is no general agreement on what sample size to use in the computation of BIC in multilevel models. The main argument for treating the number of level-3 units as sample size is that these are the independent sources of information. In this example, however, conclusions do not change if the number of level-2 units is used as sample size for BIC.

Model I is the model without random effects, while the others contain level-2 and/or level-3 random intercepts. In the parametric (normal) specifications, the integrals in the log-likelihood function were approximated using ten quadrature nodes per dimension. In order to verify the stability of the results, the models were also estimated with many more than ten quadrature point as well as with the GLLAMM

Table 1. Fit measures for the estimated models

Model	Level-2	Level-3	Log-likelihood	#parameters	BIC
I	no	no	−2188.38	7	4404.68
II	normal	no	−1711.76	8	3455.43
III	no	normal	−2061.09	8	4158.08
IV	normal	normal	−1708.72	9	3453.34
V	2-class	no	−1754.67	9	3545.24
VI	3-class	no	−1697.42	11	3438.72
VII	4-class	no	−1689.47	13	3430.80
VIII	5-class	no	−1686.02	15	3431.87
IX	no	2-class	−2092.24	9	4220.38
X	no	3-class	−2058.09	11	4160.06
XI	no	4-class	−2053.77	13	4159.40
XII	no	5-class	−2053.76	15	4167.35
XIII	4-class	2-class	−1687.85	15	3435.53

adaptive quadrature option (Rabe-Hesketh, Skrondal and Pickles, 2002). For the model with two random effects, the log-likelihood stayed more or less the same. For the models with a single random effect, we obtained somewhat higher log-likelihood values. With 50 quadrature points, Models II and III gave log-likelihood values of −1710.46 and −2058.23, respectively.

In order to give an impression about computation time, estimation of the most extended parametric model (Model IV) took less than 20 seconds with the experimental version of Latent GOLD, while estimation with GLLAMM took about ten times as long. It should be noted that GLLAMM is not only slower because it uses Newton–Raphson with numerical derivatives, but also because it is written in an interpreter language (STATA). Estimation of any of the nonparametric models with our code took less than a second. Although this option is not documented, GLLAMM can also be used to estimate nonparametric models with more than two levels (Rabe-Hesketh, personal communication).

The fit measures of the reported models show that the level-2 variance is clearly significant (compare Model II and Models V–VIII with Model I). The higher log-likelihood values and the lower BIC values indicate that the nonparametric models (Models VI–VIII) capture the heterogeneity in the intercept somewhat better than the parametric model (Model II). Based on the BIC values of Models V–VIII, it can be concluded that in the nonparametric approach no more than four latent classes of level-2 units are needed.

If we do not take into account the level-2 variation in the intercept, there is also clear evidence for a level-3 effect on the intercept (compare Model III and Models IX–XII with Model I). On the other hand, if the level-2 variation is taken into account, the importance of the level-3 variation reduces enormously: In terms of BIC, Model IV is only slightly better than Model II and Model XIII is even worse than Model VII. What is clear from the test results is that the between individuals (level-2) variation is much more important than the between districts (level-3) variation.

Table 2. Parameter estimates for models I, II, IV, VI, and XIII

	Model I	Model II	Model IV	Model VII	Model XIII
Fixed effects					
Intercept	1.50 (0.07)	1.97 (0.13)	2.09 (0.18)	0.97 (0.16)	0.96 (0.17)
Time					
1983	−0.13 (0.08)	−0.16 (0.08)	−0.16 (0.08)	−0.16 (0.08)	−0.16 (0.08)
1984	−0.55 (0.07)	−0.68 (0.08)	−0.68 (0.08)	−0.67 (0.08)	−0.67 (0.08)
1985	−0.22 (0.08)	−0.27 (0.08)	−0.27 (0.08)	−0.26 (0.08)	−0.26 (0.08)
Religion					
Catholic	−1.08 (0.10)	−1.07 (0.21)	−1.59 (0.32)	−1.64 (0.25)	−1.32 (0.32)
Protestant	−0.38 (0.06)	−0.49 (0.19)	−0.71 (0.21)	−0.22 (0.14)	−0.29 (0.16)
Other	−0.82 (0.08)	−1.12 (0.17)	−1.32 (0.24)	−0.66 (0.17)	−0.78 (0.20)
Random intercepts					
Level-2 standard deviation		1.20 (0.05)	1.21 (0.07)	1.43	1.38
Level-3 standard deviation			0.47 (0.33)		0.28
Class-sizes					
Level-2, $t = 1$				0.33 (0.05)	0.42 (0.05)
Level-2, $t = 2$				0.29 (0.04)	0.34 (0.05)
Level-2, $t = 3$				0.21 (0.03)	0.22 (0.04)
Level-2, $t = 4$				0.17 (0.06)	0.02 (0.03)
Level-3, $m = 1$					0.62 (0.20)
Level-3, $m = 2$					0.38 (0.20)
Class-specific intercepts					
Level-2, $t = 2$				1.16 (0.12)	1.26 (0.13)
Level-2, $t = 3$				3.39 (0.27)	3.52 (0.26)
Level-2, $t = 4$				−0.77 (0.11)	−1.19 (0.50)
Level-3, $m = 2$					−0.57 (0.14)

Table 2 reports the parameter estimates for Models I, II, IV, VII, and XIII. As far as the fixed part is concerned, the substantive conclusions would be similar in all five models. The attitudes are most positive at the last time point (reference category) and most negative at the second time point. Furthermore, the effects of religion show that people without religion (reference category) are most in favor and Roman Catholics and Others are most against abortion. Protestants have a position that is close to the no-religion group.

A natural manner to quantify the importance of the random intercept terms is by their contribution to the total variance. The level-1 variance can be set equal to the variance of the logistic distribution ($\pi^2/3 = 3.29$), yielding a total variance equal to $3.29 + 1.21^2 + 0.47^2 = 4.98$, in Model IV. Thus, after controlling for the time and religion effects, in Model IV, the level-2 and level-3 variances equal 29% ($1.21^2/4.98$) and 4% ($0.47^2/4.98$) of the total variance, respectively.

The random part of the latent class regression models can be interpreted in two different ways. On the one hand, we can name the latent classes based on their coefficients. Note that the parameters for the first class are fixed to zero for identification, which amounts to using dummy coding with class 1 as reference category. On the other hand, using basic statistics calculus, one can compute the level-2 and level-3 standard deviations from the class sizes and class-specific regression coefficients, which are the parameters of the random part of the model in the parametric approach. In Model XIII, the level-2 standard deviation equals 1.38, which is somewhat higher than in the parametric model, and the level-3 standard deviation equals 0.28, which is lower than in the parametric model. These numbers correspond with variance contributions of 36 and 1 percent, respectively.

5 Discussion

An EM algorithm was presented for the ML estimation of hierarchical nonlinear models. This upward–downward method prevents the need for processing the full posterior distribution, which becomes infeasible with more than a few level-2 units per level-3 unit. The relevant marginal posterior distributions can be obtained by making use of the conditional independence assumptions underlying the hierarchical model. As was shown, it is straightforward to generalize the method to models with more than three levels.

A limitation of the parametric approach is that the numerical integration to be performed for parameter estimation can involve summation over a large number of quadrature points when the number of random effects is increased. Despite the fact that the number of points per dimension can be somewhat reduced with multiple random effects and adaptive quadrature, computational burden becomes enormous with more than five or six random coefficients. There exist other methods for computing high-dimensional integrals, like Bayesian simulation and simulated likelihood methods, but these are also computationally intensive.

As indicated by Vermunt and Van Dijk (2001), these practical problems do not occur when using a nonparametric random-effects model since the sum appearing in the log-likelihood function will always be over a small number of latent classes. For instance, computation time was less than a second for each of the latent class regression models presented in Table 1, and computation time does not increase very much if also the three time effects are assumed to be random (class specific). The nonparametric approach is not only attractive for this reason, but also because it does not rely on strong unverifiable assumptions about the random effects (Aitkin, 1999). In certain situations, one may prefer to use such strong distributional assumptions, for instance, because they yield a more parsimonious description of the heterogeneity. In the application, however, we saw that the latent class model captured much better the level-2 variation than the model with normally distributed random effects. This is a clear indication that the parametric specification is too restrictive for this data set.

References

Aitkin, M. (1999), A general maximum likelihood analysis of variance components in generalized linear models, *Biometrics* **55**, 218–234.

Agresti, A., J. G. Booth, J. P. Hobert and B. Caffo (2000), Random-effects modeling of categorical response data, *Sociological Methodology* **30**, 27–80.

BAUM, L. E., T. PETRIE, G. SOULES and N. WEISS (1970), A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *Annals of Mathematical Statistics* **41**, 164–171.

BOCK, R. D. and M. AIKIN (1981), Marginal maximum likelihood estimation of item parameters, *Psychometrika* **46**, 443–459.

DEMPSTER, A. P., N. M. LAIRD and D. R. RUBIN (1977), Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society Ser. B.* **39**, 1–38.

GOLDSTEIN, H. (1995), *Multilevel statistical models*, Halsted Press, New York.

HEDEKER, D. and R. D. GIBBONS (1996), MIXOR: A computer program for mixed effects ordinal regression analysis, *Computer Methods and Programs in Biomedicine* **49**, 157–176.

JUANG, B. H. and L. R. RABINER (1991), Hidden Markov models for speech recognition, *Technometrics* **33**, 251–272.

LAIRD, N. (1978), Nonparametric maximum likelihood estimation of a mixture distribution, *Journal of the American Statistical Association* **73**, 805–811.

LESAFFRE, E. and B. SPIESSENS (2001), On the effect of the number of quadrature points in a logistic random-effects model: an example, *Applied Statistics* **50**, 325–335.

MCGRATH, K. and J. WATERTON (1986), *British social attitudes, 1983–1986 panel survey*, Social and Community Planning Research, Technical Report, London.

RABE-HESKETH, S., A. PICKLES and A. SKRONDAL (2001), GLLAMM: A general class of multilevel models and a Stata program, *Multilevel Modelling Newsletter* **13**, 17–23.

RABE-HESKETH, S., A. SKRONDAL and A. PICKLES (2002), Reliable estimation of generalised linear mixed models using adaptive quadrature, *The Stata Journal* **2**, 1–21.

RABE-HESKETH, S., A. SKRONDAL and A. PICKLES (2003), Generalized multilevel structural equation modelling, *Psychometrika*, in press.

STROUD, A. H. and D. SECREST (1966), *Gaussian Quadrature Formulas*, Prentice Hall, Englewood Cliffs, NJ.

VERMUNT, J. K. (2003), Multilevel latent class models, *Sociological Methodology* **33**, to appear.

VERMUNT, J. K. and L. VAN DIJK (2001), A nonparametric random-coefficients approach: the latent class regression model, *Multilevel Modelling Newsletter* **13**, 6–13.

VERMUNT, J. K. and J. MAGIDSON (2000), *Latent GOLD 2.0 User's Guide*, Statistical Innovations Inc., Belmont, MA.

WEDEL, M. and W. DESARBO (2002), Mixture regression models, in: J. HAGENAARS and A. MCCUTCHEON (eds), *Applied Latent Class Analysis*, Cambridge University Press, Cambridge, 366–382.

Received: April 2003. Revised: September 2003.