Pairwise likelihood for the longitudinal mixed Rasch model

M.-L. Feddag\textsuperscript{a,∗}, S. Bacci\textsuperscript{b}

\textsuperscript{a} Department of Statistics, University of Warwick, Coventry CV4 7AL, UK
\textsuperscript{b} Department of Statistics, University of Florence Viale Morgagni, 59 50134, Italy

A R T I C L E  I N F O

Article history:
Received 26 July 2007
Received in revised form 24 August 2008
Accepted 28 August 2008
Available online 2 September 2008

A B S T R A C T

Inference in Generalized linear mixed models with multivariate random effects is often made cumbersome by the high-dimensional intractable integrals involved in the marginal likelihood. An inferential methodology based on the marginal pairwise likelihood approach is proposed. This method belonging to the broad class of composite likelihood involves marginal pairs probabilities of the responses which has analytical expression for the probit version of the model, from where we derived those of the logit version. The different results are illustrated with a simulation study and with an analysis of a real data from health-related quality of life.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

Generalized linear mixed models (GLMMs) are extensions of generalized linear models (GLMs) which accommodate correlated and overdispersed data by adding random effects to the linear predictor. Their broad applications are useful in various disciplines, such as the analysis of clustered data including longitudinal data or repeated measures. Such generalized linear mixed models are also increasingly used in various fields where subjective variables need to be measured using questionnaires with dichotomous or polytomous items. This is usual in health sciences and clinical trials, where those subjective variables might be for example pain, depression or quality of life. In such fields, the Rasch model is the most popular item response theory (IRT) model ([Rasch, 1960, 1961] and [Fischer and Molenaar, 1995]). It belongs to the family of logistic linear mixed models where the fixed effects parameters and the random effects are called the difficulty item parameters and latent traits or person parameters, respectively (Fischer and Molenaar, 1995, p. 39–51).

In the case of quality of life in clinical trials, the appearance of repeated measures arises when the same set of items is presented to a sample of persons at different time points under different conditions. More complex Rasch models have received attention in the literature. Multidimensional mixed Rasch models with item parameters considered equal across time points are analyzed by Hoijtink (1995) and Feddag and Mesbah (2005). Another kind of models which generalizes the multidimensional Rasch model is considered by Adams et al. (1997).

A problem inherent in such models is that the marginal likelihood function obtained after integrating over the multidimensional random effects always involves intractable integrals. Therefore, several kinds of approximation are considered. One approach consists in numerical approximation of the integral (Bock and Aitkin, 1981; Rigdon and Tsutakawa, 1983; Hedeker and Gibbons, 1994; Pinheiro and Bates, 1995; Meng and Schilling, 1996), so that the marginal likelihood can be computed and optimized. In Bayesian inference, Naylor and Smith (1982) have used adaptive Gaussian quadrature for the approximation of the posteriors. However, these techniques become increasingly more difficult when the dimension of the random effects increases. The second class of methods is to approximate the integrand, so that the integral of the approximation has closed form. These methods are generally based on a series expansion of the conditional
mean or the likelihood about 0 or about the current estimates of the random effects. This class includes the penalized quasi-likelihood (PQL) and the marginal quasi-likelihood (MQL) methods defined by Breslow and Clayton (1993) and Breslow and Lin (1995), and the GEE approach proposed by Feddag and Mesbah (2005). The main disadvantage of these methods which are based on the Taylor approximations is that they are accurate only for small variance components. There are three other methods which have been investigated for bias correction in GLMM with binary data by Ng et al. (2006): The first one known as bootstrap bias correction proposed by Kuk (1995) and implemented in MLwiN (Rasbash et al., 2000), the second is the Robbins–Monro (RM) stochastic approximation method (Wetherill and Glazebrook, 1986) and the last one is the simulated maximum likelihood (SML) (McCulloch, 1997). The simulation study shows that SML performs as efficiently as the other two methods and also yields standard errors of the bias-corrected parameter estimates.

Alternatively to these classical approaches, we propose the pairwise likelihood method (denoted by PL) to estimate simultaneously the fixed effect parameter and the variance component of the Rasch model. In contrast with the approximate two methods and also yields standard error of the bias-corrected parameter estimates. The main feature of a pseudo-likelihood function is that it is composed of pieces of likelihoods, a fact which can be exploited in order to prove general results about the consistency and asymptotic normality of pseudo-likelihood estimators (Lindsay, 1988).


Considering the Rasch models with fixed effects, there is previous work based on the conditional pairwise likelihood, in which the person parameters are eliminated. Zwinderman (1995) has proposed a pairwise parameter estimation in Rasch models where the item parameters are estimated by the pairwise likelihood of the all pairs of item responses given the latent trait. The obtained estimators are consistent and similar in efficiency to conditional maximum likelihood (CML) and maximum marginal likelihood (MML) estimators. For the Rasch model with ordered response categories, Andrich and Luo (2003) proposed a pairwise conditional algorithm by the use of principal components. This method of estimation has two main advantages: consistency and adaptation for missing data. For a structural probit model with binary measurements, Muthén (1984) and Muthén and Satorra (1995) have proposed a three-stage approach to obtain estimates of the different parameters, the standard errors, and a chi-squared measure of fit. The first two steps are related to the estimation of the parameters of an unstructured multivariate regression. In the first stage, the parameters (thresholds or intercepts and regression coefficients) are obtained by maximizing the individual univariate first order likelihood functions. The second step consists of the estimation of the correlation or covariances by the marginal pairwise likelihood given the parameter values estimated in the first step. We note that in both stages, the marginal first order and pairwise probabilities are not multiplied together but maximized individually. A weighted least squares (WLS) estimator is used in the last stage to fit the structural model from the asymptotic covariance matrix of the statistics obtained in the first two steps.

However, pairwise likelihood estimation has not yet been studied in the framework of the longitudinal Rasch model. In this contribution the potential of using the pairwise likelihood approach in the context of this model is evaluated and its performance is compared with the MML approach, which is more traditionally used where the integrals are approximated by Gauss–Hermite quadrature. We point out that this proposed marginal approach is different from the one proposed by Andrich and Luo (2003), which is based on conditional pairwise likelihood and from the Muthén and Satorra (1995) approach which is based on three steps.

The specific outline of the paper is as follows. We present in Section 2 the studied model with logit link. Section 3 is devoted to the existing methods of estimation for the model. In Section 4, we give the definition of the marginal pairwise likelihood and some details on the marginal pairs of probabilities. We present thereafter in Section 5 some simulation results in which our proposed approach is compared to the MML one where the integrals are approximated by Gauss Hermite quadrature. On the real data from a health-related quality of life, these two approaches are illustrated and compared to those obtained by the use of Conquest (Wu et al., 2005), GLLAMM (Rabe-Hesketh et al., 2001) and WinBUGS (Spiegelhalter et al., 2003) software. A discussion is finally presented in Section 6.

2. The model

In our study, we consider binary responses of a questionnaire which is administered to the same subjects at various occasions \( t, t = 1, \ldots, T \). In such case, the responses are correlated in two ways. Firstly, at a given time point, the binary responses of a single individual are correlated and secondly, when repeated, they become also longitudinally correlated.

From now on, we will consider a sample of \( N \) independent \((T \times 1)\) random multivariate binary observations \( Y_t = (Y_t^1, \ldots, Y_t^J)' \), \( i = 1, \ldots, N \), where \( J \) is the number of items. The vector \( Y_t^i = (Y_{t1}^i, \ldots, Y_{tJ}^i)' \) is the response vector of individual \( i \) to the questionnaire at time \( t \) and \( Y_{tj}^i \) is the binary variable response of individual \( i \) to item \( j \) at time \( t \) \(( t = 1, \ldots, T \)). Let \( Y = (Y_1, \ldots, Y_N) \) be the vector of the variables, \( b_i \) the random effect associated with subject \( i \) at time \( t \) and \( b_{tij} = (b_{t1i}, \ldots, b_{tJi})' \) be the multidimensional random effect for subject \( i \). We denote by \( y \) a realization of the random variable \( Y \). The longitudinal mixed Rasch model which has been considered by Feddag and Mesbah (2005), satisfies the following assumptions:
Given the random effect $b_{it}$, $i = 1, \ldots, N$, we have

$$P \left( Y_{ij}^T = y_{ij}^T, \ldots, Y_{ij} = y_{ij} \mid b_{ij}, \beta \right) = \prod_{t=1}^{T} \prod_{j=1}^{J} P(Y_{ij}^T = y_{ij}^T \mid b_{it}, \beta_j),$$

(1)

where $\beta_j$ and $b_{it}$ are the fixed effects parameter associated to item $j$ and the random effects associated to the subject $i$ at time $t$, respectively.

- For all $i, j, t; i = 1, \ldots, N, j = 1, \ldots, J, t = 1, \ldots, T$, the probability distribution of the random variable $Y_{ij}^T$ is given by

$$P \left( Y_{ij}^T = y_{ij}^T \mid b_{it}, \beta_j \right) = \frac{\exp \left( (b_{it} - \beta_j)y_{ij}^T \right)}{1 + \exp(b_{it} - \beta_j)}.$$

(2)

- The linear predictor $\eta_{ijt} = (b_{it} - \beta_j)$ is related to the expectation of $Y_{ij}^T$ through the link function $g$ by

$$g \left( E(Y_{ij}^T) \right) = \eta_{ijt}.$$

where $g$ is the link function, which could be the logit defined by logit$(\pi) = \ln(\pi / (1 - \pi))$ or the probit $\Phi^{-1}(\cdot)$, where $\Phi(.)$ is the cumulative distribution function (CDF) of the normal variable with mean 0 and variance 1.

- The random effects $b_1, \ldots, b_N$, are independent and identically normally distributed with mean vector $\mu = (\mu_t)_{t=1,\ldots,T}$ and covariance matrix $\Sigma = (\sigma_{ij})_{i,j=1,\ldots,T}$.

The model as formulated above is not identifiable, so some suitable restrictions have to be imposed. The classical constraint we made on the parameters is $\sum_{j=1}^{J} \beta_j = 0$.

We are interested in estimating the fixed effects parameters $\beta = (\beta_1, \ldots, \beta_J)$, the mean $\mu = (\mu_1, \ldots, \mu_T)$, the vector of the variances $\alpha = (\sigma_{11}, \sigma_{22}, \ldots, \sigma_{TT})$, and the vector of the covariances $\gamma = (\sigma_{12}, \ldots, \sigma_{1T}, \sigma_{23}, \ldots, \sigma_{T-1, T})$.

The marginal likelihood of $y$ is then given by

$$L(\beta, \mu, \alpha, \gamma \mid y) = \prod_{i=1}^{N} \int_{\mathbb{R}^T} \prod_{t=1}^{T} \prod_{j=1}^{J} \exp \left\{ (b_{it} - \beta_j)y_{ij}^T \right\} \frac{1}{1 + \exp(b_{it} - \beta_j)}/ \varphi(b_i, \mu, \alpha, \gamma) \, db_i,$$

(3)

where

$$\varphi(b_i, \mu, \alpha, \gamma) = \frac{1}{(2\pi)^{T/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (b_i - \mu) \Sigma^{-1} (b_i - \mu) \right\}$$

is the multivariate normal distribution with mean vector 0 and covariance matrix $\Sigma(\alpha, \gamma)$.

By the use of the change of variable $b_i^* = b_i - \mu_i$, the expression (3) is equivalent to the following:

$$L(\beta, \mu, \alpha, \gamma \mid y) = \prod_{i=1}^{N} \int_{\mathbb{R}^T} \prod_{t=1}^{T} \prod_{j=1}^{J} \exp \left\{ (b_{it}^* + \mu_t - \beta_j)y_{ij}^T \right\} \frac{1}{1 + \exp(b_{it}^* + \mu_t - \beta_j)}/ \varphi(b_i^*, \mu, \alpha, \gamma) \, db_i^*,$$

(4)

where $\varphi(b_i^*, \mu, \alpha, \gamma)$ is the density function of the centered vector $b_i^*$.

The maximization of this marginal likelihood is computationally difficult and needs numerical evaluation of $T$-dimensional integrals. The main approaches used to maximize (4) are described in the next section. Feddag and Mesbah (2005) have proposed the GEE approach based on Taylor approximation of the conditional likelihood which is valid only for small variance components. As an alternative to these methods, we propose the marginal pairwise likelihood approach which reduces the dimension of the integrals.

### 3. Classical methods

There are several methods for maximizing the marginal likelihood (3), which can be classified as direct and indirect maximization. The most common are the Expectation-Maximization (EM) algorithm (Dempster et al., 1977), Newton–Raphson or Fisher scoring algorithms and Quasi-Newton methods. These maximization approaches use the integration methods described below to estimate the different parameters. There are two main methods to approximate the intractable integral, and they can be classified as follows: the first one is deterministic (i.e. non-stochastic) and the second is Monte Carlo based (i.e. stochastic) numerical approximation.

Let us write the expression (4), more succinctly as

$$L(\beta, \mu, \alpha, \gamma \mid y) = \prod_{i=1}^{N} \int_{\mathbb{R}^T} h(y_i \mid \beta, \mu, b_i^*) \varphi(b_i^*, \mu, \alpha, \gamma) \, db_i^*$$

$$= \prod_{i=1}^{N} I_i(h).$$

(5)
3.1. Numerical integration

In each maximization method, the integral is replaced by a finite sum and then maximized. We begin by discussing Gaussian quadrature and then consider adaptive quadrature.

3.1.1. Gaussian quadrature

When the random effects are assumed to be normally distributed, the non-stochastic numerical approximation is usually done via Gauss–Hermite (GH) or Gauss–Legendre (GL) quadrature (Rigdon and Tsutakawa, 1983). With these quadrature methods, unidimensional integrals of the form \( \int_{t_i}^{t_f} w(t) f(t) dt \) are approximated by \( \sum_{i=1}^{m} f(t_i) v_i \), where \( t_i \) and \( v_i \) are the Gauss quadrature nodes and weights that can be found in Abramowitz and Stegun (1974) or could be computed in the statmod package of the R software (R Development Core Team, 2005). We use GH quadrature when \( w(t) = \exp(-t^2/2) \) and \( (a, b) = (-\infty, +\infty) \), and GL quadrature when \( w(t) = 1 \) and \( (a, b) = (-1, +1) \).

Integration over a multivariate normal distribution of dimension \( T \) is an extension of the unidimensional case. We have used the GH quadrature by the standardizing transformation in the form \( b^* = Cz \), where \( C \) is the lower triangular Cholesky factor for \( \Sigma (\Sigma = CC^t) \) and \( Z \) is the standardized multivariate normally distributed random variable. Then the transformed version of the integral involved in (5) takes the form

\[
l_i(h) = |C| \int_{\mathbb{R}^T} h(y_i | \beta, \mu, Cz_i) \varphi(Cz_i, \alpha, \gamma) dZ_i = \frac{1}{(2\pi)^{T/2}} \int_{\mathbb{R}^T} h(y_i | \beta, \mu, Cz_i) \exp \left(-\frac{1}{2}Z_i^t \Sigma Z_i \right) dZ_i.
\]

Hence this integral is approximated by

\[
l_i(m) = \frac{1}{(2\pi)^{T/2}} \sum_{t_1}^{m} \ldots \sum_{t_T}^{m} w_{t_1} \ldots w_{t_T} h \left( y_i | \beta, \mu, C d_{t_1, \ldots, t_T} \right),
\]

where \( d_{t_1, \ldots, t_T} = (d_{t_1} \ldots d_{t_T})^t \) and \( w_{t_1}, \ldots, w_{t_T} \) are GH quadrature nodes and weights, respectively. The accuracy of the approximation depends on the number of nodes \( m \) and on the proximity of \( h(.) \) to a polynomial of degree \( 2m - 1 \). For none of the GLMMs is \( h(y_i | \beta, \mu, b^*_t) \) a polynomial, and moreover it remains unclear what degree a well-approximating polynomial should have. Hence, there is no agreement in the literature concerning the number of nodes. As the number of dimensions \( T \) increases, the computational burden rapidly grows because the total number \( (m^T) \) of nodes increases at an exponential rate. The approximation can also be poor for large random-effects variances. Thus, for high-dimensional problems, adaptive quadrature or stochastic approximation may be a resolution.

3.1.2. Adaptive quadrature

To overcome the problems with ordinary quadrature, adaptive quadrature essentially shifts and scales the quadrature locations to place them under the peak of the integrand. For Bayesian methods, Naylor and Smith (1965) have used adaptive quadrature and importance sampling techniques to estimate the posterior distribution. It has been also considered by Pinheiro and Bates (1995) for two levels non-linear mixed models (NLMM) and by Rabe-Hesketh et al. (2005) for the multilevel models as implemented in GLLAMM (Rabe-Hesketh et al., 2001). A general overview of these estimation methods is given in Skrondal and Rabe-Hesketh (2004, Ch. 6, pp. 159–215).

3.2. Monte Carlo integration

3.2.1. Ordinary Monte Carlo integration

As an alternative to the non-stochastic Gaussian quadratures rules is Monte Carlo integration. The classical method we discuss here simulate the likelihood rather than really computing it. The integral \( l_i(h) \) can be viewed as an expectation of the function \( h(y_i | \beta, \mu, b^*_t) \) of a normally distributed random variable \( b^*_t \):

\[
\mathbb{E}[h(y_i | \beta, \mu, b^*_t)] = \int_{\mathbb{R}^T} h(y_i | \beta, \mu, b^*_t) \varphi(b^*_t, \alpha, \gamma) db^*_t.
\]

A straightforward finite-sample approximation to each expectation \( \mathbb{E}[h(y_i | \beta, \mu, b^*_t)] \) is calculated by sampling \( m \) independent realizations, \( b^*_{t1}, \ldots, b^*_{tm} \), from the multivariate normal distribution of \( b^*_t \) and then computing the sample average

\[
\frac{1}{m} \sum_{i=1}^{m} h(y_i | \beta, \mu, b^*_{t_i}).
\]

As \( m \) approaches infinity, the sample average converges to the true expectation. If both \( N \) and \( m \) go to infinity, the maximum likelihood estimators converge to their true value under suitable regularity conditions.
3.2.2. Importance sampling

Ordinary Monte Carlo integration discussed above can be improved by using importance sampling to reduce the sampling variance. A convenient chosen importance density \( g(b^*_i) \) is used to simulate \( \mathbb{E} \left[ h(y_i \mid \beta, \mu, b^*_i) \right] \) when it is either difficult to sample \( b^*_i \) from \( \psi(b^*_i; \alpha, \gamma) \) or \( \psi(b^*_i; \alpha, \gamma) \) is not smooth. The integral is then written as

\[
\mathbb{E} \left[ h(y_i \mid \beta, \mu, b^*_i) \right] = \int_{\mathbb{R}^2} g(b^*_i) \frac{h(y_i \mid \beta, \mu, b^*_i) \psi(b^*_i; \alpha, \gamma)}{g(b^*_i)} \, db^*_i.
\]

Then the integral is approximated by

\[
\mathbb{E} \left[ h(y_i \mid \beta, \mu, b^*_i) \right] \approx \frac{1}{m} \sum_{i=1}^{m} \frac{h(y_i \mid \beta, \mu, b^*_i) \psi(b^*_i; \alpha, \gamma)}{g(b^*_i)},
\]

where \( b^*_1, \ldots, b^*_m \) are sampled from the density \( g(.) \). This approach has been used in non-linear mixed effects models by Pinheiro and Bates (1995). It is analogous to adaptive quadrature, which can be viewed as a deterministic version of importance sampling.

3.3. Optimization methods and software

Once the problem of the intractable integral is solved, the maximization of the marginal likelihood as function of the fixed effects and variance components can commence. Algorithms for solving this optimization problem differ in the information they extract from the log-likelihood to find the maximum, so we can distinguish two classes: direct and indirect maximization algorithms.

The direct maximization algorithms uses only function values to locate the maximum. The simplex algorithm (Nelder and Mead, 1965) and Newton–Raphson or Fisher scoring algorithms are probably the best-known examples of this class. There are several popular software packages that aim at a direct maximization of the marginal GLMM likelihood: these include PROC NLMIXED (SAS Institute, Inc., 1999), GLLAMM (Rabe-Hesketh et al., 2001). With both of these packages, one can choose between adaptive and non-adaptive Gaussian quadrature, which could be deterministic or non-deterministic for PROC NLMIXED and only deterministic for GLLAMM. This last package uses Stata’s maximum likelihood functions to maximize the marginal likelihood by means of a modified Newton–Raphson algorithm based on numerical first and second derivatives of the marginal likelihood.

The major indirect maximization algorithm is the expectation-maximization (EM) algorithm. The E-step consists in the calculation of the expected value of the complete data log-likelihood given the current estimates of the fixed effects \( \beta, \mu \), the variance components \( (\gamma, \alpha) \) and the observed data. This quantity is approximated by Gauss quadrature or Monte Carlo integration. The M-step maximizes the function calculated in the E-step. These two steps are iterated until convergence. This algorithm with the integrals approximated by Gauss quadrature or Monte Carlo methods, is traditionally favoured over more direct maximization of an approximation to the expression (3) (see Hoijtink (1995) and Adams et al. (1997)). This algorithm for estimating IRT-type GLMMs is implemented in Conquest software (Wu et al., 2005), by means of the Newton–Raphson algorithm in the M step.

4. Pairwise likelihood approach

Alternatively to the classical approaches presented in the introduction, say MML or those based on Taylor approximation, we propose the pairwise likelihood method to estimate simultaneously the fixed effect parameters and the variance component of the described model. In contrast to some of the approximate methods, this approach does not have any restriction on the variance component.

Let denote \( \theta = (\beta', \mu', \alpha', \gamma') \) the parameters of the model. The pairwise likelihood which takes into account all the marginal pairs of probabilities within and between times is given by

\[
L_2(\theta; y) = \prod_{i=1}^{N} L_{2,i}(\theta; y_i)
\]

\[
= \prod_{i=1}^{N} \prod_{S_{i,1}} P \left( \bar{Y}_{ij}^t = \bar{y}_{ij}^t, Y_{ilj}^t = y_{ilj}^t \right) \prod_{S_{i,2}} P \left( Y_{ij}^t = y_{ij}^t, Y_{il}^h = y_{il}^h \right),
\]

where the sets \( S_{i,1} \) and \( S_{i,2} \) are, respectively, given by

\[
S_{i,1} = \{t, j, b; t = 1, \ldots, T, 1 \leq j < l \leq J \},
\]

\[
S_{i,2} = \{t, h, j, l; 1 \leq t < h \leq T, 1 \leq j, l \leq J \}.
\]

These marginal pairwise probabilities are straightforward to calculate in terms of univariate and bivariate probit for the multilevel probit models as suggested by Renard et al. (2004).
For the logit version of the model, the marginal pairwise probabilities are not straightforward. There are derived from the expression of the logistic distribution function as a normal scale mixture (see Monahan and Stefanski (1989) and Drum and McCullagh (1993)), which is given by

\[
F(t) = \frac{e^t}{1 + e^t} \simeq \sum_{i=1}^{k} p_{k,i} \Phi(tS_{k,i}),
\]

where \((S_{k,i}, p_{k,i})\) are tabled for \(k = 1, \ldots, 8\). This approximation is a compromise between Gauss Hermite quadrature the method proposed by Crouch and Spiegelman (1990). It is well accurate for \(k\) as small as 3 and improves with \(k = 5\).

Then the two first marginal probabilities are approximated as follows:

\[
P(Y_{i}^t = 1; \theta) = \sum_{i=1}^{k} p_{k,i} \Phi \left( \frac{(\mu_t - \beta_j)S_{k,i}}{\sqrt{1 + \sigma_t S_{k,i}^2}} \right),
\]

\[
P(Y_{i}^t = 1, Y_{i}^{m'} = 1; \theta) = \sum_{i=1}^{k} \sum_{h=1}^{k} p_{k,i} p_{k,h} \Phi_2 \left( \frac{(\mu_t - \beta_j)S_{k,i}}{\sqrt{1 + \sigma_t S_{k,i}^2}}, \frac{(\mu_m - \beta_j')S_{k,h}}{\sqrt{1 + \sigma_m S_{k,h}^2}}, \rho^t(k, h) \right),
\]

\[
P(Y_{i}^t = 1, Y_{i}^{m'} = 1; \theta) = \sum_{i=1}^{k} \sum_{h=1}^{k} p_{k,i} p_{k,h} \Phi_2 \left( \frac{(\mu_t - \beta_j)S_{k,i}}{\sqrt{1 + \sigma_t S_{k,i}^2}}, \frac{(\mu_m - \beta_j')S_{k,h}}{\sqrt{1 + \sigma_m S_{k,h}^2}}, \rho^{(t,m)}(k, h) \right),
\]

where

\[
\rho^t(k, h) = \frac{S_{k,i} S_{k,h} \sigma_t}{\sqrt{1 + \sigma_t S_{k,i}^2} \sqrt{1 + \sigma_t S_{k,h}^2}},
\]

\[
\rho^{(t,m)}(k, h) = \frac{S_{k,i} S_{k,h} \sigma_m}{\sqrt{1 + \sigma_t S_{k,i}^2} \sqrt{1 + \sigma_m S_{k,h}^2}},
\]

and the functions \(\Phi\) and \(\Phi_2(x; y; \rho)\) denote, respectively, the standardized normal distribution and the standardized bivariate normal distribution with correlation coefficient \(\rho\).

The other marginal pair probabilities corresponding to the combinations \((1, 0)\), \((0, 1)\) and \((0, 0)\) are straightforward by using the above probabilities.

The PL estimators \(\hat{\theta}_{MPL}\) are obtained by maximizing the whole log PL function given by

\[
l_2(\theta; y) = \sum_{i=1}^{N} l_{2,i}(\theta; y_i) = \sum_{i=1}^{N} \log L_{2,i}(\theta; y_i). \tag{8}
\]

It follows from the standard theory of estimating equations (e.g., Lindsay (1988) and Cox and Reid (2004)) that as the number of individuals \(N\) increases, \(\hat{\theta}_{MPL}\) is asymptotically normal with mean \(\theta\) and asymptotic covariance matrix \(J^{-1}K^{-1}\), where

\[
J = \mathbb{E} \left( -\sum_{i=1}^{N} \frac{\partial^2}{\partial \theta \partial \theta^T} l_{2,i} \right),
\]

and

\[
K = \mathbb{E} \left( \sum_{i=1}^{N} \frac{\partial}{\partial \theta} l_{2,i} \frac{\partial}{\partial \theta^T} l_{2,i} \right).
\]

To obtain sample estimates of the standard errors of \(\hat{\theta}_{MPL}\), we estimate \(J\) and \(K\) by

\[
\hat{J} = -\sum_{i=1}^{N} \left\{ \sum_{S_{1,i}} \frac{\partial}{\partial \theta} \log p_{i,j}^{t'} \frac{\partial}{\partial \theta^T} \log p_{i,j}^{r'} \sum_{S_{2,i}} \frac{\partial}{\partial \theta} \log p_{i,j}^{h} \frac{\partial}{\partial \theta^T} \log p_{i,j}^{h} \right\},
\]

\[
\hat{K} = \sum_{i=1}^{N} \left\{ \frac{\partial}{\partial \theta} l_{2,i}(\hat{\theta}) \frac{\partial}{\partial \theta^T} l_{2,i}(\hat{\theta}) \right\}, \tag{9}
\]

where \(p_{i,j}^{t'} = P \left( y_{i}^{t'}, y_{i}^{r'}; \hat{\theta} \right), p_{i,j}^{h} = P \left( y_{i}^{t'}, y_{i}^{h}; \hat{\theta} \right)\)
Table 1
Simulation results with parameters $\beta$, $\mu$, $\alpha$ and $\gamma$ for $N = 100$

<table>
<thead>
<tr>
<th>Param.</th>
<th>True</th>
<th>PL</th>
<th>MML$^{10}_{GH}$</th>
<th>MML$^{20}_{GH}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>S.d.</td>
<td>Mean</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>$-1$</td>
<td>$-1.023$</td>
<td>.129</td>
<td>$-1.015$</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>$-0.5$</td>
<td>$-0.498$</td>
<td>.129</td>
<td>$-0.497$</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>$0.5$</td>
<td>$0.511$</td>
<td>.128</td>
<td>$0.509$</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>$1$</td>
<td>$1.010$</td>
<td>.125</td>
<td>$1.004$</td>
</tr>
<tr>
<td>$\mu_1$</td>
<td>$1$</td>
<td>$1.040$</td>
<td>.162</td>
<td>$1.037$</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>$0$</td>
<td>$0.034$</td>
<td>.174</td>
<td>$0.032$</td>
</tr>
<tr>
<td>$\mu_3$</td>
<td>$-1$</td>
<td>$-0.991$</td>
<td>.168</td>
<td>$-0.998$</td>
</tr>
<tr>
<td>$\sigma_{11}$</td>
<td>$1$</td>
<td>$1.107$</td>
<td>.556</td>
<td>$1.084$</td>
</tr>
<tr>
<td>$\sigma_{22}$</td>
<td>$1.5$</td>
<td>$1.603$</td>
<td>.691</td>
<td>$1.527$</td>
</tr>
<tr>
<td>$\sigma_{33}$</td>
<td>$1$</td>
<td>$1.025$</td>
<td>.552</td>
<td>$1.033$</td>
</tr>
<tr>
<td>$\sigma_{12}$</td>
<td>$0.6$</td>
<td>$0.629$</td>
<td>.321</td>
<td>$0.625$</td>
</tr>
<tr>
<td>$\sigma_{13}$</td>
<td>$0.5$</td>
<td>$0.501$</td>
<td>.321</td>
<td>$0.492$</td>
</tr>
<tr>
<td>$\sigma_{23}$</td>
<td>$0.3$</td>
<td>$0.321$</td>
<td>.322</td>
<td>$0.322$</td>
</tr>
</tbody>
</table>

Table 2
Simulation results with parameters $\beta$, $\mu$, $\alpha$ and $\gamma$ for $N = 300$

<table>
<thead>
<tr>
<th>Param.</th>
<th>True</th>
<th>PL</th>
<th>MML$^{10}_{GH}$</th>
<th>MML$^{20}_{GH}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>S.d.</td>
<td>Mean</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>$-1$</td>
<td>$-0.994$</td>
<td>.078</td>
<td>$-0.994$</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>$-0.5$</td>
<td>$-0.508$</td>
<td>.070</td>
<td>$-0.507$</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>$0.5$</td>
<td>$0.505$</td>
<td>.078</td>
<td>$0.505$</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>$1$</td>
<td>$1.005$</td>
<td>.072</td>
<td>$0.996$</td>
</tr>
<tr>
<td>$\mu_1$</td>
<td>$1$</td>
<td>$1.012$</td>
<td>.104</td>
<td>$0.992$</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>$0$</td>
<td>$0.003$</td>
<td>.094</td>
<td>$0.002$</td>
</tr>
<tr>
<td>$\mu_3$</td>
<td>$-1$</td>
<td>$-1.012$</td>
<td>.102</td>
<td>$-1.015$</td>
</tr>
<tr>
<td>$\sigma_{11}$</td>
<td>$1.5$</td>
<td>$1.551$</td>
<td>.394</td>
<td>$1.529$</td>
</tr>
<tr>
<td>$\sigma_{22}$</td>
<td>$1.5$</td>
<td>$1.007$</td>
<td>.318</td>
<td>$1.019$</td>
</tr>
<tr>
<td>$\sigma_{12}$</td>
<td>$0.6$</td>
<td>$0.606$</td>
<td>.186</td>
<td>$0.608$</td>
</tr>
<tr>
<td>$\sigma_{13}$</td>
<td>$0.5$</td>
<td>$0.505$</td>
<td>.160</td>
<td>$0.508$</td>
</tr>
<tr>
<td>$\sigma_{23}$</td>
<td>$0.3$</td>
<td>$0.307$</td>
<td>.182</td>
<td>$0.309$</td>
</tr>
</tbody>
</table>

5. Illustrations

This approach is illustrated by a simulation study to evaluate its performance and is compared with the MML approach (denoted by MML$^{m}_{GH}$) where the integrals are approximated by GH quadrature with $m$ points ($m = 10, 20$). On the real data it is further compared with MML by use of Conquest which uses the EM algorithm, by use of GLLAMM software which approximates the integrals by adaptive quadrature; and with the Bayesian version of the model using WinBUGS software. We note that the Conquest software provides standard errors only for the fixed effects (item difficulty and mean of the latent trait) estimates.

5.1. Simulation study

In this section, we use simulation study to evaluate the sample performance of our proposed PL and to compare with MML using Gauss–Hermite quadrature with 10 and 20 points. The optimization method used for these two approaches is the quasi-Newton method implemented in the optim function of the R software (R Development Core Team, 2005). We simulated 200 datasets of the model for two sample sizes $N = 100, 300$, four items ($J = 4$), three repetitions ($T = 3$) and one set of parameters fixed as:

$\beta = (-1, -0.5, 0.5, 1)$,
$\mu = (1, 0, -1)$,
$\Sigma$ with $\alpha = (1, 1.5, 1)$ and $\gamma = (0.6, 0.5, 0.3)$.

We give in Tables 1 and 2 the mean and the standard deviation (s.d.) of the estimates values obtained for each size based on 200 replications.

We note that for both sample sizes, even if MML$^{m}_{GH}$ seems more accurate with $m = 20$ than with $m = 10$, there is no significant difference between their results: for almost all the parameters, the bias is the same and the standard deviation is a bit larger with 10 points.

Table 1, corresponding to the first size $N = 100$, shows that the bias for all the parameter estimates given by both approaches is very small. However, as expected the results for the size 300 given in Table 2, are better: the bias is smaller (negligible) and the standard deviation decreases. With regard to the standard deviations: for almost all the parameters, standard deviations...
for the $MML^m_{GH}$ approach with both $m = 10, 20$, are smaller than those for the PL method. That indicates as expected the loss of efficiency with the PL approach.

In terms of running time (mn denotes minutes), the PL approach is much faster than the $MML^m_{GH}$ one. For example, for each data set with $N = 100$, the PL takes 2.18 mn to converge whereas the $MML^m_{GH}$ needs 2.59 mn with $m = 10$ and 21.96 mn with $m = 20$. With the second size $N = 300$, PL reached convergence in 7.05 mn, $MML^{GH}$ in 7.83 mn and $MML^m_{GH}$ in 48.67 mn. So, as expected, the difference in running time between the two approaches is more considerable with $m = 20$ than $m = 10$.

5.2. Terminal cancer patients data

The case study described in this contribution concerns the measurement of Health-related Quality of Life (HrQoL) in cancer patients. The study is referred to a survey carried out by a national multicentric study called “Staging” (and coordinated by Dr. Franco Toscani) in 58 Italian palliative care centers in the years 1995–1996. Analyzed data are referred to terminal or advanced cancer patients undergoing palliative care.

The main purpose of HrQoL measurement is the acquisition of useful information to calibrate the palliative care. Indeed, whereas in other kinds of therapies the main objective is the survival of the patient and the HrQoL is a secondary (necessarily submitted) aim, in terminal patients, unfortunately, the survival can not be any more pursued: pain and palliative therapies aim mainly to alleviate the last life period of sick persons and, therefore, HrQoL acquires a special meaning. Because of the latent nature of the HrQoL, a measurement model is necessary: the properties of specific objectivity and uni-dimensionality (Rasch, 1960) suggest the Rasch-type models as particularly suitable.

485 people answered to the so-called Therapy Impact Questionnaire (TIQ) at baseline, i.e. immediately before the beginning of treatment. The same questionnaire was again submitted after the beginning of the palliative therapy once a week until the death or, in most fortunate cases, until the end of the study period. Because of the considerably different survival times of patients, the number of completed questionnaires is very variable: minimum is 1 (only survey at baseline) and maximum is 32 (survey at baseline and 31 weeks under palliative care). As default, questionnaires with at least 6 missing values and/or referring to patients in coma were classified as invalid.

The Therapy Impact Questionnaire (Tamburini et al., 1992) is a cancer-specific instrument to measure HrQoL and it was realized in 1987 at Pain Therapy and Palliative Care division of National Cancer Institute in Milano (Italy). The questionnaire is composed of 36 items assessing the degree of discomfort experienced by cancer patients on a verbal Likert scale with four possible answers (“not at all” = high HrQoL, “slight”, “a lot”, “very much” = low HrQoL); the period of reference is a week. The questionnaire is composed of two sections: the first one collects data on 24 physical symptoms identified through a previous evaluation of 2000 patients: only symptoms reported in more than 1% of the records were selected. A group of 12 symptoms follows, concerning psychological impact of illness on the HrQoL, and more specifically any difficulties in performing normal daily activities (difficulty at work or doing the housework; difficulty with usual free-time activities; required help to eat, get dressed or go to the toilet), emotional status (felt sad or depressed; felt anxious or scared; felt nervous, restless, irritable; felt unsure), cognitive factors (difficulty concentrating or paying attention; difficulty relaxing) and the degree of interaction with the surrounding environment (arguments with the family; felt isolated).

In the TIQ questionnaire two main latent dimensions can be detected according to a Rasch-type model: a Psychological one and a Physical one. The former is related to the mental sphere of patient, while the latter is related to physical symptoms and side effects of illness:

- Psychological or impact therapy dimension. It is composed of 8 items: difficulty in performing usual free time activities (Item 1), fatigue (Item 2), illness feeling (Item 3), sad or depressed feeling (Item 4), difficulty in concentrating or paying attention (Item 5), nervousness feeling (Item 6), insecurity (Item 7) and confusion feeling (Item 8).
- Physical dimension. It is composed by 13 items: weakness, pain, dry mouth, constipation, problems sleeping, difficulty in breathing, drowsiness, coughing, stomachache, sweating, tremors, headache and dizziness.

In order to develop the case study of estimation based on pairwise likelihood, a subset of 253 patients with 3 consecutive measurements of HrQoL is considered. Moreover, the analysis is carried out only on the psychological dimension: for a future development of the work it could be interesting to take into account also a multidimensional longitudinal Rasch model. The responses are dichotomized: modality “not at all” (= 0) versus modalities “slight”, “a lot”, “very much” (= 1). The former one means the absence of some disorder or symptom, i.e. it means a good level of HrQoL; the latter one means the presence (at a some level) of some disorder or symptom, i.e. it means a worse level of HrQoL.

Table 3 presents the estimation of the different parameters — including the correlation coefficients — and their standard errors (s.e.) obtained by the two previous approaches (PL and $MML^m_{GH}$), the Conquest software (denoted by $MML^m_{EM}$), the GLLAMM software (denoted by $GLLAMM$) and the Bayesian version of the model with MCMC by the use of WinBUGS software (denoted by Gibbs). We note that for both $MML^m_{GH}$ and $MML^m_{EM}$ methods, we have used twenty adaptive quadrature points and with GLLAMM, the number is chosen by default (eight points).

For the Bayesian version of the model, the estimates are obtained by Gibbs sampling with the following priors on the parameters:

- $\beta = (\beta_1, \ldots, \beta_7) \sim N_7(0, V)$ with $V^{-1} = \text{diag}(10^{-5})$
- $\mu \sim N_3(0, A)$ with $A^{-1} = \text{diag}(10^{-5})$
Table 3
Parameter estimates for (β, μ, α, γ), their standard errors (s.e) and correlation coefficients of the variance components

<table>
<thead>
<tr>
<th>Par.</th>
<th>PL</th>
<th>MML</th>
<th>MML</th>
<th>GLLAM</th>
<th>Gibbs(Γ1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>β1</td>
<td>−1.64</td>
<td>.13</td>
<td>−1.63</td>
<td>.11</td>
<td>−1.63</td>
</tr>
<tr>
<td>β2</td>
<td>1.77</td>
<td>.14</td>
<td>1.77</td>
<td>.12</td>
<td>1.77</td>
</tr>
<tr>
<td>β3</td>
<td>0.28</td>
<td>.13</td>
<td>0.28</td>
<td>.11</td>
<td>0.28</td>
</tr>
<tr>
<td>β4</td>
<td>1.21</td>
<td>.14</td>
<td>1.21</td>
<td>.12</td>
<td>1.20</td>
</tr>
<tr>
<td>β5</td>
<td>0.26</td>
<td>.11</td>
<td>0.26</td>
<td>.11</td>
<td>0.27</td>
</tr>
<tr>
<td>β6</td>
<td>−0.46</td>
<td>.12</td>
<td>−0.45</td>
<td>.11</td>
<td>−0.45</td>
</tr>
<tr>
<td>β7</td>
<td>−0.98</td>
<td>.12</td>
<td>−0.98</td>
<td>.11</td>
<td>−0.98</td>
</tr>
<tr>
<td>β8</td>
<td>−0.45</td>
<td>*</td>
<td>−0.45</td>
<td>*</td>
<td>−0.45</td>
</tr>
<tr>
<td>µ1</td>
<td>−1.06</td>
<td>.12</td>
<td>−1.09</td>
<td>.09</td>
<td>−1.08</td>
</tr>
<tr>
<td>µ2</td>
<td>−0.99</td>
<td>.11</td>
<td>−1.04</td>
<td>.09</td>
<td>−1.03</td>
</tr>
<tr>
<td>µ3</td>
<td>−1.26</td>
<td>.14</td>
<td>−1.23</td>
<td>.10</td>
<td>−1.24</td>
</tr>
<tr>
<td>σ11</td>
<td>1.75</td>
<td>.30</td>
<td>1.89</td>
<td>*</td>
<td>1.85</td>
</tr>
<tr>
<td>σ33</td>
<td>1.81</td>
<td>.33</td>
<td>2.07</td>
<td>*</td>
<td>2.04</td>
</tr>
<tr>
<td>σ13</td>
<td>2.77</td>
<td>.42</td>
<td>2.41</td>
<td>*</td>
<td>2.46</td>
</tr>
<tr>
<td>σ12</td>
<td>1.73</td>
<td>.33</td>
<td>1.94</td>
<td>*</td>
<td>1.88</td>
</tr>
<tr>
<td>σ23</td>
<td>1.65</td>
<td>.30</td>
<td>1.76</td>
<td>*</td>
<td>1.71</td>
</tr>
<tr>
<td>ρ12</td>
<td>1.98</td>
<td>.31</td>
<td>2.00</td>
<td>*</td>
<td>2.06</td>
</tr>
<tr>
<td>ρ13</td>
<td>0.97</td>
<td>*</td>
<td>0.98</td>
<td>*</td>
<td>0.97</td>
</tr>
<tr>
<td>ρ23</td>
<td>0.75</td>
<td>*</td>
<td>0.83</td>
<td>*</td>
<td>0.80</td>
</tr>
<tr>
<td>ρ33</td>
<td>0.88</td>
<td>*</td>
<td>0.90</td>
<td>*</td>
<td>0.92</td>
</tr>
<tr>
<td>Running time</td>
<td>54 mn</td>
<td>63 mn</td>
<td>2 h 33 mn</td>
<td>6 h 40 mn</td>
<td>9 h 50 mn</td>
</tr>
</tbody>
</table>

- $b_i \sim N_3(0, W^{-1})$, where $W \sim Wishart(I^*, 3)$
- two different matrices $I^*$:

$$I_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad I_2 = \begin{pmatrix} 2 & .7 & .7 \\ .7 & 2 & .7 \\ .7 & .7 & 1.9 \end{pmatrix}.$$

The results obtained with the two matrices $I_1$ and $I_2$ are similar, so we have reported only those corresponding to $I_1$. These results are based on 20,000 samples where the length of burn-in (number of iterations which are discarded at the beginning) is 2000. There is an example in the manual help of WinBUGS for the Rasch model with univariate latent trait to the Law school aptitude test (LSAT) data, where the inference is based on 10,000 samples and 1000 burn-in.

From Table 3, we note that there are not relevant variations among estimates for the fixed effects parameters under the different methods. However, there is difference for the variance component estimates, where some estimates differ by more than 15%. Some more evident difference can be observed for the standard error estimates, which seem to be larger for the PL approach: this result indicates as expected the loss of efficiency with this approach relative to the others which use the full likelihood. However, the differences are not as such to preclude the same interpretation of the outputs.

As regards the difficulty item estimates, we can see that Item 1 (Confusion) is the "easiest" one: that is, the presence of confusion is very common in all patients, also in patients with a good level of HrQoL. More precisely, it is the only item with a probability greater than 50.0% to show itself for a patient with an average level of HrQoL, independently of the time point. By applying the Formula dichotomous Rasch model, one can calculate the probability that a patient with a level of HrQoL equal to $-1$ (i.e. the approximate average HrQoL level of the cancer patients’ population, in any time point and according to all the methods) shows confusion: it is equal to 65.2%. On the contrary, the most “difficult” item is Item 2 (Feeling tired): only patients with a very low level of HrQoL have an high probability to present this symptom. In particular, the probability that a patient with a level of HrQoL equal to $-1$ feels tired is only 5.9%.

As regards the measures of HrQoL, the estimated average values of the cancer patients’ population are very similar over time (independently of the method); only for time 3 we can observe a slight improvement: In fact, due to the coding used for the items, a high score of the estimates means a low level of HrQoL, whereas a low score means a high level of HrQoL. This result can be explained by taking into account that two opposite effects act on the HrQoL in terminal patients: on one side the negative effect of approaching death, and on the other side the positive effect of the palliative care. In order to separate these two effects we need a control group of terminal cancer patients who are not subject to palliative care or who are subjected to different types of care: unfortunately, this is not our case. We can conclude that the average level of HrQoL is constant over the time. Obviously, it does not exclude a different trend of quality of life for some patients (for details see Bacci (2008)). We can also observe that the levels of HrQoL are very strongly and positively correlated, although the correlation is decreasing over time (indeed, $p_{13}$ is always smaller than $p_{12}$ and $p_{23}$).

Finally, despite the different optimization methods used in these approaches (quasi–Newton for PL and MML, modified Newton–Raphson for GLLAMM and Newton–Raphson for Conquest), we give the running time for each of them. The PL approach is the fastest one: it takes only 54 min, a running time comparable only to the MML (63 min), whereas the other three methods all take much more time.
6. Discussion

This paper describes a pairwise likelihood approach to estimate simultaneously the fixed effects parameters (item difficulty and mean of the latent traits) and the variance components of the longitudinal Rasch model with logit link. This approach which belongs to the broad class of pseudo-likelihood is developed to solve the estimation problems of the Rasch model in the longitudinal case, where in general, the multidimensional integral involved in the marginal likelihood is intractable. It provides consistent and asymptotically normal estimators and the simulation study shows that the bias of the estimates is almost zero. This proposed method performs well and the loss of efficiency relative to maximum marginal likelihood seems to be not substantial as shown by the simulation study. However, in order to know how good or bad the standard errors are, it will be interesting even it is too demanding to provide in the simulation study the mean standard errors for comparison with the standard deviations. In the accuracy point of view, it is difficult to assess how large the discrepancy might be in a given application. This is in contrast to adaptive quadrature methods where the only approximation is to use numerical integration. In this case, the accuracy of estimates can be assessed by increasing the number of quadrature points. In terms of running time, the simulation study shows that the PL approach is much faster than the MML with Gauss Hermite quadrature approximation using either ten and twenty points. For larger value of \( T \), the reduction of running time provided by the PL approach will be more substantial. Further, this proposed approach could be easily generalized to the two-parameter logistic model with discrimination parameters.

Acknowledgements

The authors would like to thank the referees for their valuable comments that led to substantive improvements of the paper. Also, the authors wish to thank Prof D. Firth whose suggestions led to greatly improve the manuscript. The first author gratefully acknowledges support from ESRC grant RES-576-25-5020, “Developing Statistical Modelling in the Social Sciences”, held jointly at the Universities of Lancaster and Warwick.

References

Ng, E.S.W., Carpenter, J.R., Goldstein, H., Rasbash, J., 2006. Estimation in generalised linear mixed models with binary outcomes by simulated maximum likelihood. Statistical Modeling 6, 23–42.


