Statistics
in Medicine

Research Article

Received 7 July 2012, Accepted 7 February 2013 Published online 11 March 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5778

Robust Bayesian inference for multivariate longitudinal data by using normal/independent distributions

Sheng Luo,[a*†] Junsheng Ma[a] and Karl D. Kieburtz[b]

Many randomized clinical trials collect multivariate longitudinal measurements in different scales, for example, binary, ordinal, and continuous. Multilevel item response models are used to evaluate the global treatment effects across multiple outcomes while accounting for all sources of correlation. Continuous measurements are often assumed to be normally distributed. But the model inference is not robust when the normality assumption is violated because of heavy tails and outliers. In this article, we develop a Bayesian method for multilevel item response models replacing the normal distributions with symmetric heavy-tailed normal/independent distributions. The inference is conducted using a Bayesian framework via Markov Chain Monte Carlo simulation implemented in `BUGS` language. Our proposed method is evaluated by simulation studies and is applied to Earlier versus Later Levodopa Therapy in Parkinson's Disease study, a motivating clinical trial assessing the effect of Levodopa therapy on the Parkinson's disease progression rate. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: item response theory; latent variable; Markov Chain Monte Carlo; robust inference; clinical trial

1. Introduction

Parkinson's disease (PD) is a chronic progressive neurodegenerative disease that is manifested clinically by tremors (trembling in hands, arms, legs, jaw, or head), rigidity (stiffness of the limbs and trunk), bradykinesia (slowness of movement), and impaired balance [1]. In the USA alone, the estimated prevalence is more than 500,000 people, and about 50,000 new cases are reported annually. Currently, the only established treatments for PD are symptomatic (therapies that only affect the disease symptoms, not the cause) [2]. Many PD clinical trials have been conducted to search for the neuroprotective treatments that are capable of halting or slowing down disease progression. In these clinical trials, patients are repeatedly measured on multiple outcomes of various types (e.g. binary, ordinal, and continuous). Hence, the multilevel data structure has three levels of nesting, that is, multiple measurements (level 1) are nested within measurement occasions (level 2) that are nested within patients (level 3). To analyze these multivariate longitudinal data, the analysis model should account for three sources of correlation, that is, inter-source (different measures at the same visit), intra-source (same measure at different visits), and cross correlation (different measures at different visits) [3].

To this end, multilevel item response theory (IRT) models (referred to as MLIRT) are often used to analyze such multivariate longitudinal data. Within the MLIRT modeling framework, the observed measurements are viewed as imperfect manifestations of the interaction between subject-specific latent traits and measurement-specific parameters (e.g. the measurement's ability to distinguish PD patients in disease severity). The latent traits are regressed on covariates of interest (e.g. treatment and disease duration) as well as confounding variables. Because the response variable in the regression model is latent rather than observed, this approach is also called latent regression [4–9]. The MLIRT models separate

[a]*Division of Biostatistics, University of Texas School of Public Health, 1200 Pressler St, Houston, Texas 77030, U.S.A.*
[b]*Department of Neurology, University of Rochester, Rochester, NY 14642, U.S.A.*
**Correspondence to: Sheng Luo, Division of Biostatistics, University of Texas School of Public Health, 1200 Pressler St, Houston, Texas 77030, U.S.A.*
†*E-mail: sheng.t.luo@uth.tmc.edu*

the measurement-specific parameters and subject-specific covariates from manifest data so that both may be understood and studied separately. Advantages of the MLIRT models include better reflection of multilevel data structure, simultaneous estimation of measurement-specific parameters and covariate effects, and accurate inference about high-level measures [10–12]. Given a distribution assumption for the latent variables, the MLIRT models are equivalent to nonlinear mixed models [13]. Marginal maximum likelihood method [8] and Bayesian method [12, 14–19] have been widely used for MLIRT model inference. For the detailed description and summary of the IRT models, please refer to Fox [17] and van der Linden and Hambleton [20].

However, when some outcomes are continuous, the analysis submodel within the IRT framework is a common factor model [21], which assumes normal random errors. Even though normality may be a reasonable model assumption, it may lack robustness in parameter estimation under departure from normality (e.g. heavy tails and outliers) [22]. Moreover, the primary efficacy evaluation in confirmatory clinical trials is often required by agencies to follow the 'intent-to-treat' (ITT) principle, that is, the analysis includes all randomized individuals regardless of the abnormal observations. By including all patients who are randomized, the ITT analysis preserves the benefits of randomization and is commonly accepted as the most unbiased approach. Hence, the potential outlying observations cannot be deleted to follow the ITT principle. Some popular data transformation methods (e.g. log, square root, Box–Cox) might generate distributions close to normality. But there are some disadvantages with transformations, for example, (i) transformation provides reduced information on an underlying data generation scheme; (ii) component-wise transformation might not guarantee joint normality; (iii) parameters might be hard to interpret on a transformed scale; and (iv) transformations may not be universal and usually vary with datasets [23]. Alternatively, the approaches based on weighting functions have been proposed to reduce the influence of response disturbances in IRT models [24, 25], whereas the approaches based on the minimum covariance determinant estimator have been used to obtain robust inference in factor analysis [26], principal component analysis [27], and discriminant analysis [28]. From a practical perspective, it is essential to replace the normal distributions with some more flexible symmetric and heavy-tailed distributions. Liu [29] discussed a class of robust distributions known as normal/independent (NI) distributions including student's t, slash, and contaminated normal distributions [30]. The NI distributions have been applied to linear regression model [31, 32], nonlinear regression model [33], linear mixed model (LME) [22, 34–37], nonlinear mixed model (NLME) [38, 39], LME and NLME with censored responses (LMEC and NLMEC) [23], joint modeling of longitudinal measurements and competing risks [40, 41], structure equation model [42], stochastic volatility models [43, 44], Grubbs' model [45], and measurement error model [46–48]. To the best of our knowledge, there is no literature on Bayesian inference for the MLIRT models using the NI distributions to relax the normality assumption for the continuous outcomes. In this article, we propose a robust Bayesian parametric method for the MLIRT models on the basis of the NI distributions and apply it to a motivating PD clinical trial.

The rest of the article is organized as follows. We describe a motivating clinical trial, the data structure, and the outlier issue in Section 2 . In Section 3 we discuss the MLIRT models, the NI distributions, the Bayesian inference, and the Bayesian model selection criteria. In Section 4 we conduct simulation studies in which the MLIRT models by using the NI distributions are compared with the MLIRT models assuming normal random errors with and without outlying measurements. In Section 5, we apply the proposed method to a motivating clinical trial dataset. Concluding remarks and discussions are given in Section 6.

2. A motivating clinical trial

This article is motivated by the ELLDOPA (Earlier versus Later Levodopa Therapy in Parkinson's Disease) study, a multicenter, placebo-controlled, randomized, dose-ranging, double-blind clinical trial conducted from year 1998 to year 2001. This study assessed the effect of levodopa (study drug) on the PD progression rate. A total of 361 patients with early PD were randomly assigned to receive levodopa at a daily dose of 150 mg (low dose, 92 patients), 300 mg (medium dose, 88 patients), and 600 mg (high dose, 91 patients), or a matching placebo (90 patients) for a period of 40 weeks. We combine the patients who received levodopa (271 patients) and refer to them as the treatment group. The details of the ELLDOPA study can be found in Fahn *et al.* [1].

The outcomes collected include QoL, unified Parkinson's disease rating scale (UPDRS) total score (referred to as UPDRS), status of fatigue (referred to as fatigue), and Schwab and England activities of daily living (referred to as SEADL), measured at four visits, that is, baseline, week 9, week 24,

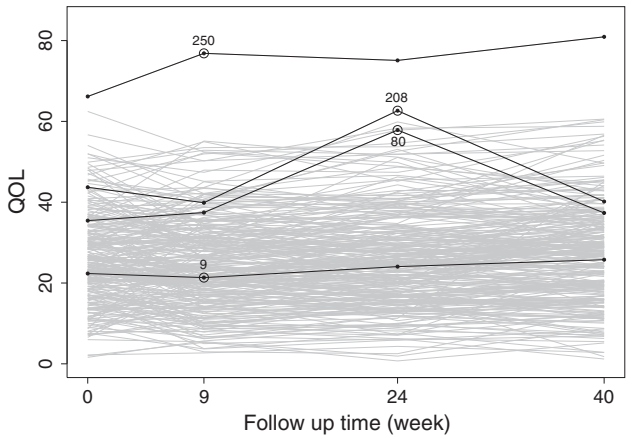

Figure 1. Longitudinal profile plots of the observed QoL measurements. Numbers 9, 80, 208, and 250 denote four patients to be used for further discussion.

and week 40. QoL is the sum of 32 questions each measured on a 5-point scale (0-4) [49]. It is rescaled to 100 so it is an approximate continuous variable with a larger value reflecting worse clinical outcomes. The UPDRS total score is the sum of 44 questions each measured on a 5-point scale (0-4), and it is approximated by a continuous variable with an integer value from 0 (not affected) to 176 (most severely affected). Fatigue is the sum of 31 questions, and it is approximated by a continuous variable with integer value from 0 (not affected) to 182 (severe) [50]. The SEADL (ordinal variable with integer value from 0 to 100 incrementing by 5, with larger value reflecting better clinical outcomes) is a measurement of activities of daily living [51]. We recode the outcome SEADL so that higher values in all outcomes are worse clinical conditions. Moreover, we combine some categories in SEADL with zero or a small number of patients so that it has 7 categories.

Figure 1 displays the longitudinal profiles of the outcome QoL. Because PD is a slow progression disease, it is unexpected to observe sudden value change in the outcome variables. Patients 80 and 208 have a change of 20.42 and 22.78 units, respectively, in the outcome QoL from week 9 visit to week 24 visit. Hence, these two measurements are potentially outlying observations. Patient 250 has much higher (worse) QoL values than all other patients. Patient 9 appears near the 'center' among all measurements. These four patients will be used later for further discussion. We are interested in investigating how the outlying measurements would affect the inference within the MLIRT modeling framework.

3. The robust item response model formulation and estimation

3.1. The multilevel item response model

Ignoring the outlying observation issue for the moment, we introduce the MLIRT model. The level 1 model describes item responses at a specific time point. The level 2 model accounts for variation in the latent traits across time within patient and between patients. Specifically, let y_{ijk} (binary, ordinal, and continuous) be the observed outcome k ($k = 1, \ldots, K$) from patient i ($i = 1, \ldots, N$) at visit j ($j = 1, \ldots, J$, where $j = 1$ is baseline). Throughout the article, we code all outcomes so that larger observation values are worse clinical conditions. Let $\boldsymbol{y}_{ij} = (y_{ij1}, \ldots, y_{ijk}, \ldots, y_{ijK})'$ be the vector of observation for patient i at visit j, and let $\boldsymbol{y}_i = (\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{iK})'$ be the outcome vector across visits. We model the binary outcomes, the cumulative probabilities of ordinal outcomes, and the continuous outcomes by using a two-parameter submodel [52], a graded response submodel [53], and a common factor submodel [21], respectively.

$$\text{logit}\left\{p(y_{ijk} = 1|\theta_{ij})\right\} = a_k + b_k\theta_{ij}, \tag{1}$$

$$\text{logit}\left\{p(y_{ijk} \leqslant l|\theta_{ij})\right\} = a_{kl} - b_k\theta_{ij}, \text{ with } l = 1, 2, \ldots, n_k - 1, \tag{2}$$

$$y_{ijk} = a_k + b_k\theta_{ij} + \epsilon_{ijk}, \tag{3}$$

where random error $\epsilon_{ijk} \sim N\left(0, \sigma_k^2\right)$ with σ_k^2 being the variance of continuous outcome k, a_k is the outcome-specific 'difficulty' parameter, and b_k is the outcome-specific 'discriminating' parameter that is always positive and represents the discrimination of outcome k, that is, the degree to which outcome k discriminates between patients with different latent disease severity θ_{ij}. In model (2), the ordinal outcome k has n_k categories and $n_k - 1$ thresholds $a_{k1}, \ldots, a_{kl}, \ldots, a_{kn_k-1}$ that must satisfy the order constraint $a_{k1} < \ldots < a_{kl} < \ldots < a_{kn_k-1}$. The probability that patient i being in category l on outcome k at visit j is $p\left(Y_{ijk} = l|\theta_{ij}\right) = p(Y_{ijk} \leqslant l|\theta_{ij}) - p(Y_{ijk} \leqslant l-1|\theta_{ij})$. The latent variable θ_{ij} is continuous and it indicates patient i's unobserved disease severity at visit j, with a higher value denoting more severe status. Models (1) to (3) consist of level 1 model that describes the item responses as functions of the outcome-specific parameters and the subject-specific latent variable at a certain visit.

At level 2, we model the disease severity θ_{ij} as a function of covariates, visit time, and random effects

$$\theta_{ij} = \boldsymbol{X}_{i0}\boldsymbol{\beta}_0 + u_{i0} + (\boldsymbol{X}_{i1}\boldsymbol{\beta}_1 + u_{i1})t_{ij}, \tag{4}$$

where $\boldsymbol{X}_{i0}$ and $\boldsymbol{X}_{i1}$ are patient i's covariate vectors including some covariates of interest (e.g. treatment assignment) and potential confounding variables (e.g. age and gender), $\boldsymbol{X}_{i0}$ and $\boldsymbol{X}_{i1}$ can share part of or all the covariates, t_{ij} is the visit time variable with $t_{i1} = 0$ for baseline, random intercept u_{i0} and random slop u_{i1} determine the subject-specific baseline disease severity and disease progression rate, respectively. The random effects u_{i0} and u_{i1} can be assumed either independent or correlated. We let $\boldsymbol{u}_i = (u_{i0}, u_{i1})'$ and assume $u_{i0} \sim \text{N}01$, $u_{i1} \sim \text{N}0\sigma_u^2$, and $\text{corr}(u_{i0}, u_{i1}) = \rho$. Model (4) is a latent trait regression model assuming that each patient has subject-specific baseline disease severity after adjusting for the covariate vector $\boldsymbol{X}_{i0}$ and that the disease severity changes linearly with subject-specific slope depending on the covariate vector $\boldsymbol{X}_{i1}$. We now give an example to further illustrate model (4). If no covariate is in $\boldsymbol{X}_{i0}$ and only the treatment assignment variable is included in $\boldsymbol{X}_{i1}$, $\theta_{ij} = u_{i0} + [\beta_{10} + \beta_{11} I_i(\text{trt}) + u_{i1}]t_{ij}$, where $I(\cdot)$ is an indicator function (1 if treatment and 0 otherwise). In this model, β_{10} and $\beta_{10} + \beta_{11}$ denote the disease progression rates for placebo and treatment patients, respectively, with β_{11} being the change in disease progression rate introduced by the treatment. The significant negative coefficient β_{11} indicates that the treatment slows down the disease progression. The combined level 1 and level 2 models are MLIRT with subject-specific covariance (referred to as subject-specific MLIRT models) [12, 14–17, 54]. The underlying assumption of linear disease progression rate in model (4) can be relaxed by adding the quadratic or higher-order term of time t, for example, $\theta_{ij} = \boldsymbol{X}_{i0}\boldsymbol{\beta}_0 + u_{i0} + (\boldsymbol{X}_{i1}\boldsymbol{\beta}_1 + u_{i1})t_{ij} + (\boldsymbol{X}_{i2}\boldsymbol{\beta}_2 + u_{i2})t_{ij}^2$, where $\boldsymbol{X}_{i0}$, $\boldsymbol{X}_{i1}$, and $\boldsymbol{X}_{i2}$ can share part of or all the covariates.

All three sources of correlations illustrated in Section 1 are accounted for via the random effect vector $\boldsymbol{u}_i$. It is well-known that item response models are overparameterized because they have more parameters than can be estimated from the data [17]. Hence, additional constraints on the location and scale of the latent disease severity are required to make models identifiable. In the subject-specific MLIRT models specified in the previous text, we establish the location and scale of the latent disease severity by setting $\text{E}u_{i0} = \text{E}u_{i1} = 0$ and $\text{Var}[u_{i0}] = 1$ so that at $t = 0$ (baseline), the disease severity θ_{ij} follows standard normal distribution.

Under the local independence assumption (i.e. conditioning on the random effect vector $\boldsymbol{u}_i$, all components in $\boldsymbol{y}_{ij}$ are independent), the full likelihood of patient i across all visits is

$$p(\boldsymbol{y}_i, \boldsymbol{u}_i) = \left[\prod_{j=1}^{J}\prod_{k=1}^{K} p(y_{ijk}|\boldsymbol{u}_i)\right] \cdot p(\boldsymbol{u}_i). \tag{5}$$

For notation convenience, we let the difficulty parameter vector be $\boldsymbol{a} = \left(\boldsymbol{a}_1', \ldots, \boldsymbol{a}_k', \ldots, \boldsymbol{a}_K'\right)'$, with $\boldsymbol{a}_k = \left(a_{k1}, \ldots, a_{kn_k-1}\right)'$, the discrimination vector be $\boldsymbol{b} = (b_1, \ldots, b_K)'$, and $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_0', \boldsymbol{\beta}_1'\right)'$, the parameter vector $\boldsymbol{\Phi}\left(\boldsymbol{a}', \boldsymbol{b}', \boldsymbol{\beta}', \rho, \sigma_u, \sigma_k\right)'$. We thereafter refer to the MLIRT model assuming normal random errors for the common factor submodel as M_1.

3.2. *Normal/idependent distributions*

In Section 3.1, we assume normal random errors for the common factor submodel, making inferences sensitive to the presence of outliers [22]. In this section, we construct the robust MLIRT models using the NI distributions.

An element of the NI family [30, 55] is defined as the distribution of the p-dimensional random vector $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{w}^{-1/2}\boldsymbol{e}$, where $\boldsymbol{\mu}$ is a location vector, $\boldsymbol{e}$ is a normally distributed random vector with mean zero and covariance matrix $\boldsymbol{\Sigma}$, $\boldsymbol{w}$ is a positive weight variable with density $p(\boldsymbol{w}|\nu)$, ν is a scalar or vector valued parameter. Given $\boldsymbol{w}$, $\boldsymbol{y}$ follows a normal distribution $N(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with the marginal density of $\boldsymbol{y}$ given by $\text{NI}(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \int p(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{w})p(\boldsymbol{w}|\nu)d\boldsymbol{w}$. The NI family provides a class of symmetric heavy-tailed distributions that consist of the multivariate version of student's t, slash, and contaminated normal distributions. When the density $p(\boldsymbol{w}|\nu)$ degenerates to $\boldsymbol{w} = 1$ (e.g. when $\nu \to \infty$), $\text{NI}(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ becomes a normal distribution as a special case.

A univariate NI distribution [30, 55], when applied to the common factor submodel (3), is $y_{ijk} = a_k + b_k\theta_{ij} + \epsilon'_{ijk}$, where $\epsilon'_{ijk} = \epsilon_{ijk}/\sqrt{w_{ijk}}$ with $\epsilon_{ijk} \sim N\left(0, \sigma_k^2\right)$, the weight w_{ijk} is a positive random variable with density $p(w_{ijk}|\nu)$, where the tuning parameter $\nu > 0$. The NI distributions provide a group of symmetric heavy-tailed distributions of ϵ'_{ijk}. In this article, we consider student's t and slash distributions. Specifically, ϵ'_{ijk} follows student's t distribution $t\left(0, \sigma_k^2, \nu\right)$, where the tuning parameter ν representing degree of freedom, when $w_{ijk} \sim \text{Gamma}(\nu/2, \nu/2)$. In addition, ϵ'_{ijk} follows slash distribution with tuning parameter ν, when $w_{ijk} \sim \text{Beta}(\nu, 1)$. Although ν in the slash distribution needs to be estimated from the data, ν in student's t distribution can be either estimated from the data or pre-specified to a small value, for example, $\nu = 3$ or 4. General principles of parsimony suggest that ν be fixed for small datasets and estimated for large ones [32]. Lange et al [32] suggests that estimated values of ν below 1 should be regarded with suspicion. When $\nu \to \infty$, the distributions $\text{Gamma}(\nu/2, \nu/2)$ and $\text{Beta}(\nu, 1)$ degenerate to 1, i.e., $w_{ijk} \equiv 1$. In this case, $\epsilon'_{ijk} = \epsilon_{ijk} \sim N\left(0, \sigma_k^2\right)$ and the NI distributions reduce to the normal distributions. In practice, the weight variable w_{ijk} can be estimated and be used for outlier detection. Specifically, if the posterior distribution of w_{ijk} has high density close to 0, it indicates that the corresponding observation can be a potential outlier [34]. Detailed examples of this outlier detection technique will be given in Section 5. For notation convenience, we let $\boldsymbol{w}_i = \{w_{ijk}, j = 1, \ldots, J, k = 1, \ldots, K\}$ and the parameter vector is $\boldsymbol{\Phi} = \left(\boldsymbol{a}', \boldsymbol{b}', \boldsymbol{\beta}', \rho, \sigma_u, \sigma_k, \nu\right)'$. The full likelihood of patient i across all visits is

$$p(\boldsymbol{y}_i, \boldsymbol{w}_i, \boldsymbol{u}_i) = \left\{\prod_{j=1}^{J}\prod_{k=1}^{K} p(y_{ijk}|\boldsymbol{u}_i, w_{ijk})p(w_{ijk})\right\} \cdot p(\boldsymbol{u}_i). \tag{6}$$

Henceforth, we refer to the MLIRT models by using the NI distributions in the common factor submodel as the NI-MLIRT models. We consider three NI-MLIRT models in this article, that is, student's t distribution with $\nu = 4$ (refer to as M_2), Student's t distribution with ν estimated (refer to as M_3), and slash distribution (refer to as M_4).

3.3. *Bayesian inference*

To infer the unknown parameter vector $\boldsymbol{\Phi}$, we use Bayesian inference based on Markov Chain Monte Carlo (MCMC) posterior simulations. We use vague priors on all elements in the parameter vector $\boldsymbol{\Phi}$. Specifically, the prior distributions of all elements in $\boldsymbol{\beta}$ are $N(0, 100)$. We use the prior distribution $b_k \sim \text{Gamma}(0.001, 0.001)$, $k = 1, \ldots, K$, to ensure positivity. The prior distribution for the difficulty parameter a_k of the continuous outcomes is $a_k \sim N(0, 10000)$, because some continuous measurements are quite large. To obtain the prior distributions for the threshold parameters of ordinal outcome k, we let a_{k1} $N(0, 100)$, and $a_{kl} = a_{k,l-1} + \delta_l$ for $l = 2, n_k - 1$, with $\delta_l \sim N(0, 100)I(0,)$, that is, normal distribution left censored at 0. We use the prior distribution $\rho \sim \text{Uniform}[-1, 1]$, and $\sigma_k, \sigma_u, \nu \sim \text{Gamma}(0.001, 0.001)$.

The model fitting is performed in `OpenBUGS` (`OpenBUGS` version 3.2.2) by specifying the likelihood function and the prior distribution of all unknown parameters. We use history plots available in `OpenBUGS` and view the absence of apparent trend in the plot as evidence of convergence. In addition, we use Gelman–Rubin diagnostic to ensure the scale reduction $\widehat{R}$ of all parameters are smaller than 1.1 [56].

3.4. *Bayesian model selection criteria*

There are a wide variety of model selection criteria in Bayesian inference. The conditional predictive ordinate (CPO) [57–60] has been widely used to assess model fit and model selection. Let $\boldsymbol{y}$ be the full

data and $\boldsymbol{y}_{(i)}$ be the data with subject i omitted. The CPO for subject i is defined as

$$\text{CPO}_i = p(\boldsymbol{y}_i|\boldsymbol{y}_{(i)}) = \int p(\boldsymbol{y}_i|\boldsymbol{\Phi})p(\boldsymbol{\Phi}|\boldsymbol{y}_{(i)})d\boldsymbol{\Phi}, \tag{7}$$

where $p(\boldsymbol{\Phi}|\boldsymbol{y}_{(i)})$ is the posterior density of $\boldsymbol{\Phi}$ given data $\boldsymbol{y}_{(i)}$. CPO is a form of cross-validation with high value indicating that the data for subject i can be accurately predicted by a model based on the data from all other subjects. Hence, a model with larger CPO_i for all subjects suggests a better fit. Although the close form of CPO_i is not available for our proposed model, a Monte Carlo estimator of CPO_i can be obtained by MCMC samples $\{\boldsymbol{\Phi}^{(t)}\}_{t=1}^{M}$ from posterior distribution $p(\boldsymbol{\Phi}|\boldsymbol{y})$, with M being the total number of post burn-in samples. Because $p(\boldsymbol{y}_i|\boldsymbol{y}_{(i)}) = p(\boldsymbol{y})/p(\boldsymbol{y}_{(i)}) = 1/\int p(\boldsymbol{\Phi}|\boldsymbol{y})/p(\boldsymbol{y}_i|\boldsymbol{y}_{(i)},\boldsymbol{\Phi})d\boldsymbol{\Phi}$, a harmonic mean approximation of CPO_i is $\widehat{\text{CPO}}_i = \left(\frac{1}{M}\sum_{t=1}^{M}\frac{1}{p(\boldsymbol{y}_i|\boldsymbol{y}_{(i)},\boldsymbol{\Phi}^{(t)})}\right)^{-1} = \left(\frac{1}{M}\sum_{t=1}^{M}\frac{1}{p(\boldsymbol{y}_i|\boldsymbol{\Phi}^{(t)})}\right)^{-1}$ [58]. A summary statistics of $\widehat{\text{CPO}}_i$ for all subjects is the log pseudo-marginal likelihood (LPML) defined as $\text{LPML} = \sum_{i=1}^{N}\log(\widehat{\text{CPO}}_i)$. A larger value of LMPL indicates better fit of the model.

Moreover, we adopt a model selection approach by using the deviance information criterion (DIC) proposed by Spiegelhalter *et al*. [61]. The DIC provides an assessment of model fitting and a penalty for model complexity. The deviance statistic is defined as $D(\boldsymbol{\Phi}) = -2\log f(\boldsymbol{y}|\boldsymbol{\Phi}) + 2\log h(\boldsymbol{y})$, where $f(\boldsymbol{y}|\boldsymbol{\Phi})$ is the likelihood function for the observed data $\boldsymbol{y}$ given the parameter vector $\boldsymbol{\Phi}$ and $h(\boldsymbol{y})$ denotes a standardizing function of the data alone that has no impact on model selection [60]. The DIC is defined as $\text{DIC} = 2\overline{D} - D(\overline{\boldsymbol{\Phi}}) = \overline{D} + p_D$, where $\overline{D} = E_{\boldsymbol{\Phi}|\boldsymbol{y}}[D]$ is the posterior mean of the deviance, $D(\overline{\boldsymbol{\Phi}}) = D(E_{\boldsymbol{\Phi}|\boldsymbol{y}}[\boldsymbol{\Phi}])$ is the deviance evaluated at the posterior mean $\overline{\boldsymbol{\Phi}}$ of the parameter vector, and $p_D = \overline{D} - D(\overline{\boldsymbol{\Phi}})$ is the effective number of parameters. A smaller value of DIC indicates a better-fitting model.

We also use the expected Akaike information criterion (EAIC) and the expected Bayesian (or Schwarz) information criterion (EBIC) as model selection tools [60]. The EAIC and EBIC can be estimated as $\text{EAIC} = \overline{D} + 2p$ and $\text{EBIC} = \overline{D} + p\log(N)$, where p is the number of elements in the parameter vector $\boldsymbol{\Phi}$. Smaller values of EAIC and EBIC indicate better fit of the model.

4. Simulation studies

In this section, we conduct three simulation studies to compare the performance of two NI-MLIRT models M_2 and M_4, and the MLIRT model M_1. The data structure of the simulated datasets is similar to the motivating ELLDOPA study and has two continuous outcomes and two ordinal outcomes with seven categories, and five visits (baseline and four follow-up visits, $J = 5$).

In the first simulation study, both continuous outcomes follow normal distributions. In the second and the third simulation studies, the first continuous outcome mostly follows a normal distribution but has 3% and 5% outliers, respectively, whereas the second continuous outcome follows a normal distribution. In all simulation studies, we generate 100 datasets with sample size $N = 400$ and no missing data. Each dataset is generated using the following algorithm.

1. Consider the treatment assignment variable x_i as the only covariate, simulate $x_i \sim \text{Bernoulli}(0.5)$.
2. Set $\boldsymbol{\beta} = (0.4, -0.5)$, $\rho = 0.4$, and $\sigma_u = 1.3$, simulate the random effects vector $\boldsymbol{u}_i \sim N_2(0, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ being a 2×2 matrix denoted by $\left((1, \rho\sigma_u), (\rho\sigma_u, \sigma_u^2)\right)$, and generate θ_{ij} for $j = 1, \ldots, J$ from model (4) with $\boldsymbol{X}_{i0} = 0$ and $\boldsymbol{X}_{i1} = x_i$.
3. To generate outliers, simulate 3% or 5% of the random errors ϵ_{ij1} from normal distributions $N(60, 100)$ and $N(-60, 100)$ with rates 30% and 70%, respectively. Set $\sigma_1 = 5$ and simulate the rest of the random errors $\epsilon_{ij1} \sim N\left(0, \sigma_1^2\right)$. Set $a_1 = 25$, $b_1 = 10$, and generate the first continuous outcome y_{ij1} from model (3).
4. Assuming no outlier in the second continuous response, set $\sigma_2 = 20$, and simulate the random errors $\epsilon_{ij2} \sim N\left(0, \sigma_2^2\right)$ for $j = 1, \ldots, J$. Set $a_2 = 80$, $b_2 = 18$, and generate the second continuous outcome y_{ij2} from model (3).
5. Set $\boldsymbol{a}_3 = (-2.7, -0.6, 2, 2.8, 5, 6)$, $b_3 = 2.0$, $\boldsymbol{a}_4 = (-0.1, 1, 1.8, 2.6, 3.3, 4)$, $b_4 = 0.4$, and simulate ordinal outcomes y_{ij3} and y_{ij4} from model (2) for $j = 1, \ldots, J$.
6. Repeat Steps 1 to 5 until the responses of all patients are generated.

Table I. True values (True), bias, standard error (SE), standard deviation (SD), and coverage probabilities (CP) of 95% credible intervals for models M_1, M_2, and M_4, when there are no outliers.

		Results for model M_1				Results for model M_2				Results for model M_4			
	True	Bias	SE	SD	CP	Bias	SE	SD	CP	Bias	SE	SD	CP
a_1	25.000	0.003	0.520	0.543	0.930	−0.057	0.524	0.521	0.920	0.001	0.520	0.504	0.930
b_1	10.000	0.014	0.378	0.412	0.910	0.075	0.383	0.398	0.930	−0.010	0.381	0.378	0.950
a_2	80.000	−0.020	1.017	1.024	0.980	−0.085	1.023	0.932	0.970	−0.006	1.016	1.013	0.950
b_2	18.000	−0.033	0.701	0.751	0.930	0.062	0.710	0.757	0.960	−0.040	0.709	0.680	0.960
a_{31}	−2.700	0.009	0.139	0.138	0.940	0.019	0.139	0.129	0.980	−0.013	0.140	0.138	0.960
a_{32}	−0.600	0.007	0.122	0.131	0.940	0.008	0.122	0.124	0.960	0.002	0.122	0.125	0.930
a_{33}	2.000	−0.004	0.132	0.145	0.920	−0.004	0.132	0.129	0.960	0.012	0.132	0.123	0.960
a_{34}	2.800	−0.008	0.141	0.149	0.950	−0.009	0.141	0.145	0.960	0.013	0.142	0.139	0.950
a_{35}	5.000	−0.017	0.186	0.183	0.930	−0.017	0.186	0.187	0.950	0.034	0.188	0.162	0.950
a_{36}	6.000	−0.004	0.213	0.220	0.930	0.000	0.213	0.216	0.960	0.052	0.216	0.194	0.970
b_3	2.000	−0.004	0.094	0.092	0.930	0.005	0.095	0.095	0.930	0.011	0.095	0.092	0.960
a_{41}	−0.100	−0.009	0.052	0.050	0.940	−0.011	0.052	0.051	0.960	0.000	0.052	0.060	0.890
a_{42}	1.000	−0.005	0.057	0.054	0.960	−0.004	0.057	0.051	0.970	−0.001	0.057	0.061	0.940
a_{43}	1.800	0.005	0.068	0.064	0.960	0.002	0.068	0.064	0.980	0.012	0.068	0.070	0.950
a_{44}	2.600	0.015	0.087	0.084	0.940	0.006	0.086	0.082	0.950	0.007	0.087	0.083	0.980
a_{45}	3.300	0.023	0.112	0.110	0.960	0.014	0.111	0.104	0.970	0.013	0.111	0.101	0.970
a_{46}	4.000	0.042	0.148	0.141	0.980	0.016	0.146	0.131	0.990	0.018	0.146	0.155	0.930
b_4	0.400	−0.002	0.027	0.026	0.930	−0.003	0.027	0.024	0.970	0.001	0.027	0.031	0.930
β_{10}	0.400	−0.005	0.091	0.095	0.940	0.001	0.091	0.098	0.960	−0.002	0.092	0.083	0.970
β_{11}	−0.500	0.010	0.124	0.137	0.910	0.008	0.123	0.132	0.910	0.002	0.126	0.120	0.970
ρ	0.400	−0.008	0.045	0.042	0.980	−0.003	0.045	0.040	0.980	−0.004	0.045	0.048	0.950
σ_u	1.300	0.003	0.064	0.060	0.970	0.004	0.064	0.067	0.940	0.006	0.065	0.058	0.970

We apply the Bayesian framework in Section 3.3 to obtain samples from the posterior distributions of the parameters of interest. For each dataset in all simulation studies, we run three parallel MCMC chains with overdispersed initial values. Each chain is run for 30,000 iterations, the first 20,000 iterations are discarded as a burn-in, and the next 10,000 samples are used to calculate the joint posterior distribution of the parameters of interest.

The results from models M_1, M_2, and M_4 of the first simulation study with no outliers are compared in Table I. In this table, we label the average of the posterior means minus the true values as bias, the square root of the average of the variances as standard error (SE), the standard deviation of the posterior means as SD, and the coverage probabilities of 95% equal-tail CI as CP. The results suggest that all three models generate comparable results, that is, the bias is negligible, SE is close to SD, and the credible interval coverage probabilities are reasonably close to nominal level of 95%.

Table II displays the results of the second simulation study with 3% of outliers in the first continuous outcome. The results from models M_2 and M_4 indicate that the estimates of all parameters have negligible bias and SE being close to SD. The coverage probabilities of 95% credible intervals are all reasonably around the nominal value. In contrast, model M_1 gives severely biased estimates, large SD and SE, and low coverage probabilities for the outcome-specific parameters a_1 and b_1, because the presence of outliers clearly violates the normality assumption for the first continuous outcome. Model M_1 provides reasonable estimates to all other parameters because the information from other response variables are sufficient to estimate other outcome-specific parameters, the regression parameter vector $\boldsymbol{\beta}$, and the random effect related parameters ρ and σ_u. Another interesting phenomena is that models M_2 and M_4 provide slightly smaller estimates of SE and SD for all parameters. Furthermore, in the presence of 5% outliers in the first continuous outcome, the bias, SE, SD, and CP of the parameter a_1 and b_1 from model M_1 further deteriorate, whereas the estimates for all other parameters are still reasonable (Table III). In comparison, models M_2 and M_4 still provide reasonable estimates for all parameters. Although the outliers do not have notable impact on the estimates and inference of the regression parameter $\boldsymbol{\beta}_1$ in Tables II and III, the biased estimates of b_1 leads to misleading clinical interpretations and conclusions, because the expected change of continuous variable y_{ij1} from baseline ($t_{i1} = 0$) to time t_{ij} is $b_1 t_{ij} \boldsymbol{X}_{i1} \boldsymbol{\beta}_1$ (i.e. $E[y_{ij1} - y_{i11}] = b_1 t_{ij} \boldsymbol{X}_{i1} \boldsymbol{\beta}_1$).

Table II. True values (True), bias, standard error (SE), standard deviation (SD), and coverage probabilities (CP) of 95% credible intervals for models M_1, M_2, and M_4, when there are 3% outliers in the first continuous response.

		Results for model M_1				Results for model M_2				Results for model M_4			
	True	Bias	SE	SD	CP	Bias	SE	SD	CP	Bias	SE	SD	CP
a_1	25.000	**−0.741**	**0.628**	**0.629**	**0.790**	−0.025	0.516	0.515	0.920	−0.010	0.519	0.514	0.910
b_1	10.000	**−0.135**	**0.448**	**0.462**	0.940	−0.014	0.377	0.382	0.920	−0.008	0.376	0.387	0.940
a_2	80.000	−0.009	1.039	1.093	0.930	−0.036	1.010	1.043	0.940	−0.024	1.013	1.033	0.940
b_2	18.000	−0.052	0.786	0.814	0.960	−0.041	0.702	0.695	0.950	−0.035	0.699	0.697	0.940
a_{31}	−2.700	−0.006	0.143	0.143	0.940	−0.012	0.140	0.146	0.960	−0.011	0.140	0.140	0.960
a_{32}	−0.600	0.005	0.124	0.128	0.930	0.008	0.122	0.128	0.930	0.004	0.122	0.126	0.910
a_{33}	2.000	0.015	0.135	0.134	0.960	0.016	0.132	0.128	0.960	0.014	0.132	0.125	0.960
a_{34}	2.800	0.016	0.145	0.151	0.940	0.021	0.141	0.143	0.940	0.017	0.142	0.141	0.950
a_{35}	5.000	0.031	0.196	0.177	0.960	0.044	0.188	0.161	0.950	0.039	0.188	0.164	0.960
a_{36}	6.000	0.044	0.226	0.202	0.970	0.063	0.216	0.195	0.960	0.057	0.216	0.197	0.950
b_3	2.000	0.010	0.103	0.096	0.950	0.016	0.095	0.091	0.950	0.013	0.094	0.093	0.970
a_{41}	−0.100	−0.002	0.052	0.060	0.890	−0.002	0.052	0.061	0.890	−0.001	0.052	0.060	0.900
a_{42}	1.000	−0.001	0.057	0.062	0.940	0.000	0.057	0.059	0.940	−0.001	0.057	0.060	0.950
a_{43}	1.800	0.013	0.068	0.072	0.950	0.011	0.068	0.069	0.960	0.012	0.068	0.069	0.950
a_{44}	2.600	0.008	0.087	0.086	0.960	0.006	0.087	0.085	0.970	0.006	0.087	0.081	0.980
a_{45}	3.300	0.019	0.112	0.106	0.960	0.017	0.111	0.101	0.970	0.012	0.111	0.099	0.980
a_{46}	4.000	0.025	0.147	0.158	0.930	0.029	0.147	0.154	0.930	0.018	0.146	0.155	0.930
b_4	0.400	0.001	0.028	0.033	0.920	0.002	0.027	0.031	0.910	0.000	0.027	0.030	0.930
β_{10}	0.400	−0.004	0.094	0.088	0.950	0.001	0.090	0.087	0.940	−0.001	0.090	0.089	0.920
β_{11}	−0.500	−0.003	0.130	0.129	0.910	0.003	0.125	0.127	0.960	−0.003	0.122	0.124	0.920
ρ	0.400	−0.001	0.050	0.054	0.930	−0.005	0.045	0.046	0.970	−0.004	0.045	0.048	0.950
σ_u	1.300	0.007	0.073	0.070	0.960	0.005	0.064	0.059	0.960	0.006	0.064	0.059	0.950

Note: Large bias, large SE and SD, and poor CP are highlighted in boldface.

From the simulation studies, we conclude that the NI-MLIRT models provide results comparable to the MLIRT model when the random errors of the continuous responses follow normal distributions, while it provides more accurate estimates for response-specific parameters and more efficient estimates for other parameters than the MLIRT model when some continuous response variable has outliers.

5. Application to the ELLDOPA study

In this section, we apply the proposed method and the Bayesian inference framework to the motivating ELLDOPA study. For all the results in this section, we use three parallel MCMC chains with overdispersed initial values, and run each chain for 50,000 iterations. The first 45,000 iterations are discarded as burn-in and the inference is based on the remaining 5,000 iterations from each chain.

To analyze the ELLDOPA dataset, we let $\boldsymbol{X}_{i0} = 0$ and consider the treatment assignment variable x_i (1 treatment, and 0 if placebo) as the only covariate in $\boldsymbol{X}_{i1}$. Hence, the level 2 model (4) is $\theta_{ij} = u_{i0} + [\beta_{10} + \beta_{11}x_i + u_{i1}]t_{ij}$, with visit times being transformed in year $t_{ij} = (0, 9, 24, 40)/52$. We first fit the MLIRT model M_1. A plot of the standardized residuals from the response QoL for all patients at each visit (Figure 2) indicates that the normal random error assumption for the response QoL does not fit very well the whole dataset. A few data points have SRs with absolute values larger than 3 (e.g. 3.969 for patient 80 at week 24 visit, 3.462 for patient 208 at week 24 visit, and 3.067 for patient 250 at week 9 visit), indicating potential outliers. In contrast, the SR for patient 9 at week 9 visit is 0.030, indicating a non-outlier. Hence, a heavy-tailed distribution for the response QoL is essential.

We then apply three NI-MLIRT models (models M_2, M_3, and M_4) to the response QoL. As pointed out in Section 3.2, the normal distribution is a special case of the NI distributions when the tuning parameter ν is large. In practice, the small estimate of ν is an indication of heavy-tailed distribution. Figure 3 displays the posterior density distributions of the degree of freedom of student's t distribution in model M_3 and the tuning parameter ν of the slash distribution in Model M_4. For both models, the densities are concentrated around small value (mean: 5.193, 95% CI: [3.386, 8.866] for student's t distribution; mean:

Table III. True values (True), bias, standard error (SE), standard deviation (SD), and coverage probabilities (CP) of 95% credible intervals for models M_1, M_2, and M_4, when there are 5% outliers in the first continuous response.

		Results for model M_1				Results for model M_2				Results for model M_4			
	True	Bias	SE	SD	CP	Bias	SE	SD	CP	Bias	SE	SD	CP
a_1	25.000	**−1.227**	**0.691**	**0.720**	**0.530**	0.054	0.519	0.568	0.900	0.072	0.525	0.527	0.910
b_1	10.000	**−0.230**	**0.470**	**0.473**	0.940	0.060	0.375	0.350	0.950	−0.007	0.373	0.391	0.940
a_2	80.000	0.064	1.043	1.188	0.910	0.140	1.014	1.113	0.930	0.108	1.025	1.109	0.900
b_2	18.000	0.054	0.801	0.825	0.930	−0.072	0.700	0.634	0.960	−0.020	0.695	0.701	0.940
a_{31}	−2.700	−0.006	0.144	0.152	0.940	−0.018	0.140	0.155	0.920	−0.018	0.141	0.141	0.950
a_{32}	−0.600	0.001	0.124	0.140	0.900	−0.001	0.122	0.139	0.920	−0.003	0.123	0.135	0.930
a_{33}	2.000	0.001	0.135	0.135	0.940	−0.001	0.132	0.134	0.970	−0.003	0.133	0.122	0.970
a_{34}	2.800	0.007	0.145	0.152	0.940	0.003	0.142	0.150	0.920	0.000	0.142	0.140	0.950
a_{35}	5.000	0.015	0.196	0.160	0.970	0.034	0.188	0.153	0.970	0.024	0.188	0.149	0.970
a_{36}	6.000	0.019	0.226	0.182	0.980	0.033	0.216	0.179	0.960	0.039	0.216	0.182	1.000
b_3	2.000	0.012	0.103	0.095	0.950	0.004	0.094	0.086	0.970	0.012	0.094	0.090	0.960
a_{41}	−0.100	−0.002	0.052	0.063	0.850	−0.005	0.052	0.057	0.900	0.001	0.052	0.062	0.890
a_{42}	1.000	−0.002	0.057	0.062	0.930	−0.006	0.057	0.061	0.920	−0.002	0.057	0.063	0.920
a_{43}	1.800	0.013	0.068	0.075	0.940	0.008	0.068	0.069	0.960	0.015	0.068	0.072	0.940
a_{44}	2.600	0.001	0.087	0.093	0.960	0.006	0.087	0.086	0.960	0.010	0.087	0.085	0.970
a_{45}	3.300	0.014	0.112	0.110	0.950	0.022	0.112	0.106	0.960	0.016	0.111	0.101	0.960
a_{46}	4.000	0.024	0.147	0.157	0.950	0.024	0.147	0.167	0.940	0.022	0.146	0.156	0.940
b_4	0.400	0.001	0.028	0.032	0.930	−0.001	0.027	0.029	0.920	0.003	0.027	0.029	0.930
β_{10}	0.400	−0.005	0.095	0.095	0.940	0.006	0.091	0.092	0.930	0.010	0.091	0.095	0.930
β_{11}	−0.500	−0.006	0.131	0.141	0.920	−0.018	0.123	0.131	0.900	−0.010	0.123	0.134	0.930
ρ	0.400	−0.002	0.050	0.053	0.950	−0.003	0.045	0.045	0.960	−0.004	0.045	0.046	0.950
σ_u	1.300	0.001	0.073	0.072	0.950	0.003	0.064	0.059	0.970	0.002	0.064	0.063	0.960

Note: Large bias, large SE and SD, and poor CP are highlighted in boldface.

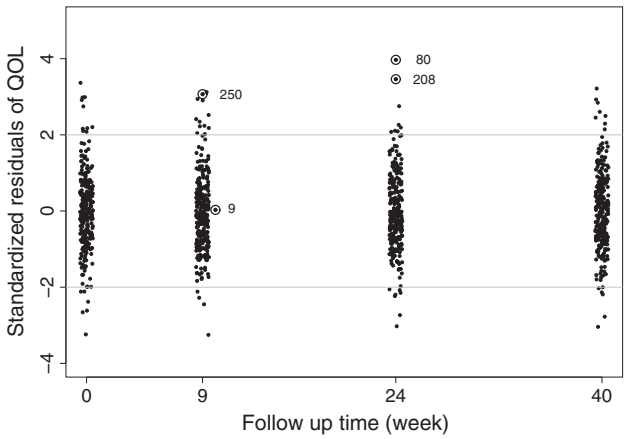

Figure 2. Standardized residuals of the response QoL for all patients at each visit. Numbers 9, 80, 208, and 250 denote four patients.

1.767, 95% CI: [1.267, 2.597] for the slash distribution) providing some evidence against the adequacy of the normality assumption made for the response QoL.

The weight variable w_{ijk} in the NI distributions can be used for outlier detection [34]. In Figure 4, the posterior distributions of the weight variable (w_{ijk}) are presented for some target patients at certain visits. Patient 208's and patient 80's QoL observations at week 24 visit are potential outliers indicated by Figure 2. Their posterior distributions of the weights are sharp with majority of the density close to zero. For patient 250 at week 9 visit, a potential outliers, the posterior distribution of weight is less sharp in two student's t distributions (models M_2 and M_3) and is quite flat in the slash distribution (model M_4).

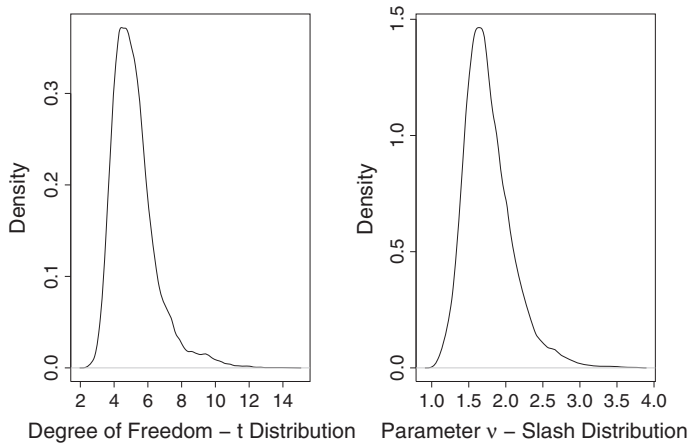

Figure 3. Posterior densities of the degree-of-freedom of student's t distribution and the tuning parameter ν of the slash distribution when applying the NI distributions to the response QoL.

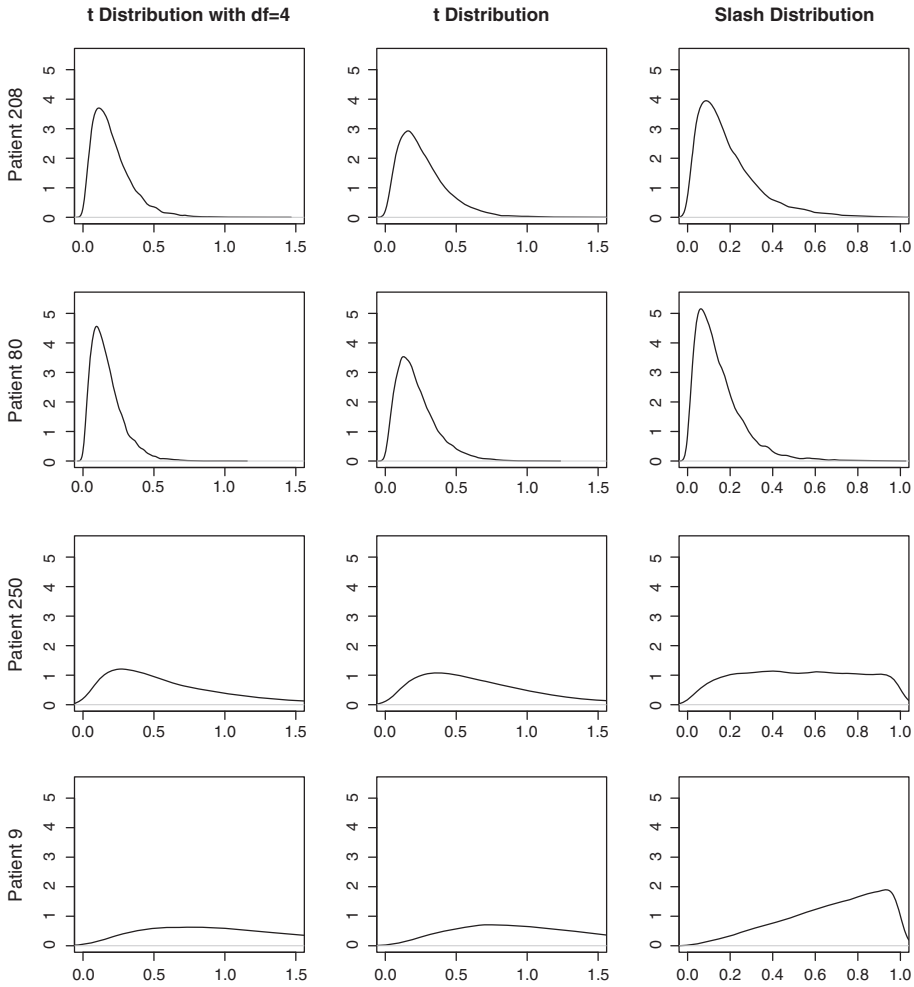

Figure 4. Estimates of the weight variable w_{ijk} for some patients from various models.

This indicates that this observation may not be an outlier because the QoL measurements at other visits are quite large (severe) as well. In contrast, for patient 9 at week 9 visit, a clear non-outlier, the posterior distributions of the weight from all three NI distributions have no density clustering at small values.

Table IV compares models M_1 with M_4 by using the model selection criteria discussed in Section 3.4. All three NI-MLIRT models perform significantly better than model M_1 with a larger LPML value and

Table IV. Results of fitting the MLIRT model M_1, three NI-MLIRT models (M_2, M_3, and M_4) to the response QoL. Parameters a_k and b_k for $k = 1, 2, 3$ are the outcome-specific parameters of the responses QoL, unified Parkinson's disease rating scale (UPDRS), and fatigue, respectively. Parameters $a_{41}, \ldots, a_{45}$ and b_4 are the outcome-specific parameters of the response SEADL.

Criterion	Model M_1			Model M_2			Model M_3			Model M_4		
LPML	−14227.0			−14145.4			−14165.4			−14171.7		
DIC	28248.0			28063.7			28083.2			28076.7		
EAIC	28001.5			27651.5			27702.1			27760.7		
EBIC	28070.0			27719.9			27774.1			27832.7		
Parameters	Mean_{SD}	95% CI		Mean_{SD}	95% CI		Mean_{SD}	95% CI		Mean_{SD}	95% CI	
a_1	$25.291_{0.645}$	24.070	26.630	$25.052_{0.596}$	23.900	26.210	$25.153_{0.626}$	23.870	26.360	$25.059_{0.635}$	23.810	26.290
b_1	$9.605_{0.473}$	8.734	10.570	$9.775_{0.475}$	8.862	10.770	$9.744_{0.485}$	8.823	10.700	$9.828_{0.497}$	8.898	10.870
a_2	$24.914_{0.515}$	23.930	25.950	$24.932_{0.477}$	24.010	25.850	$24.988_{0.490}$	24.020	25.940	$24.907_{0.499}$	23.910	25.870
b_2	$6.418_{0.438}$	5.588	7.309	$6.180_{0.442}$	5.360	7.100	$6.183_{0.438}$	5.343	7.062	$6.213_{0.440}$	5.369	7.102
a_3	$82.523_{1.484}$	79.700	85.500	$82.520_{1.400}$	79.770	85.260	$82.672_{1.423}$	79.850	85.420	$82.390_{1.438}$	79.530	85.190
b_3	$18.760_{1.208}$	16.490	21.270	$18.651_{1.184}$	16.400	21.060	$18.634_{1.219}$	16.350	21.110	$18.794_{1.227}$	16.510	21.270
a_{41}	$-2.256_{0.143}$	−2.543	−1.987	$-2.201_{0.127}$	−2.451	−1.955	$-2.213_{0.127}$	−2.466	−1.963	$-2.195_{0.134}$	−2.458	−1.932
a_{42}	$-0.032_{0.114}$	−0.263	0.189	$-0.036_{0.103}$	−0.238	0.166	$-0.046_{0.105}$	−0.252	0.164	$-0.031_{0.108}$	−0.245	0.182
a_{43}	$2.435_{0.148}$	2.162	2.736	$2.355_{0.141}$	2.086	2.637	$2.350_{0.143}$	2.081	2.635	$2.365_{0.142}$	2.096	2.652
a_{44}	$3.168_{0.167}$	2.849	3.506	$3.073_{0.161}$	2.765	3.396	$3.067_{0.164}$	2.756	3.395	$3.085_{0.162}$	2.779	3.415
a_{45}	$4.777_{0.243}$	4.319	5.266	$4.658_{0.236}$	4.212	5.132	$4.653_{0.239}$	4.195	5.140	$4.669_{0.240}$	4.225	5.150
b_4	$1.387_{0.105}$	1.190	1.602	$1.298_{0.101}$	1.112	1.508	$1.300_{0.102}$	1.107	1.507	$1.310_{0.102}$	1.119	1.517
β_{10}	$1.015_{0.134}$	0.757	1.287	$0.935_{0.129}$	0.686	1.201	$0.945_{0.130}$	0.698	1.212	$0.942_{0.128}$	0.696	1.201
β_{11}	$-0.892_{0.145}$	−1.185	−0.612	$-0.817_{0.141}$	−1.107	−0.549	$-0.826_{0.141}$	−1.112	−0.557	$-0.817_{0.139}$	−1.097	−0.553
ρ	$0.630_{0.160}$	0.258	0.853	$0.653_{0.153}$	0.292	0.864	$0.649_{0.167}$	0.310	0.921	$0.660_{0.164}$	0.312	0.887
σ_u	$0.258_{0.058}$	0.160	0.379	$0.273_{0.061}$	0.172	0.419	$0.276_{0.062}$	0.179	0.428	$0.268_{0.059}$	0.164	0.398

LPML, log pseudo-marginal likelihood; DIC, deviance information criterion; EAIC, expected Akaike information criterion; EBIC, expected Bayesian (or Schwarz) information criterion.

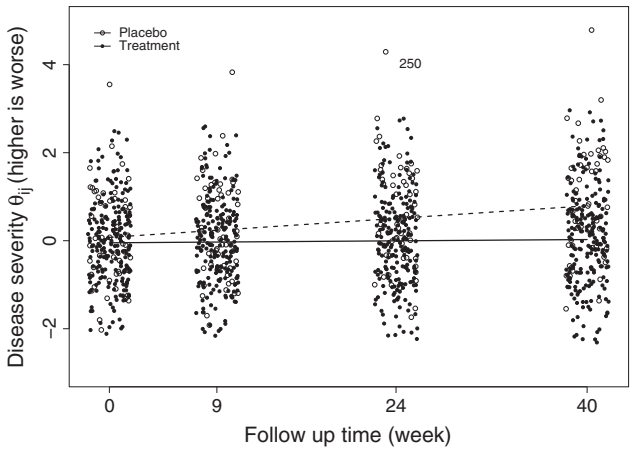

Figure 5. Estimates of the subject-specific disease severity θ_{ij} at each visit and the lowess smooth curves for the two groups.

smaller DIC, EAIC, and EBIC values, suggesting the necessity of accounting for the outliers in the response QoL. Model M_2 has the best fit with the highest LPML value and the lowest DIC, EAIC, and EBIC values. Model M_2 provides better fit than model M_3, because the fixed value of ν at 4 in model M_2 is relatively close to the estimate of ν (mean: 5.193 and 95% CI: [3.386, 8.866]) from model M_3. The more parsimonious model M_2 provides better fit in this scenario. Table IV compares the posterior mean, SD, 95% equal-tail CIs from various models. The results from all three NI-MLIRT models are very close to each other. In contrast, Model M_1 tends to give larger parameter estimates (especially β_{10} and β_{11}) and it is less precise with larger SDs and wider CIs, a phenomena also reported in Rosa *et al.* [34] and Lachos et al [23].

The results from model M_2 in Table IV indicate that the placebo patients show significant deterioration across time with the disease progression rate being 0.935 units per year ($\widehat{\beta_{10}}$, 95% CI: [0.686, 1.201]). Although the treatment patients also show significant deterioration across time with disease progression rate being 0.118 units per year ($\widehat{\beta}_{10} + \widehat{\beta}_{11}$, 95% CI: [0.006, 0.230]), the treatment significantly slows down the disease progression rate by −0.817 ($\widehat{\beta}_{11}$, 95% CI: [−1.107, −0.549]) units per year, suggesting the efficacy of the study drug levodopa. To visualize the difference in the disease progression rates in the two groups, Figure 5 displays the posterior estimate of each patient's subject-specific latent disease severity at each visit. The lowess smooth curves [62] for the placebo and the treatment groups are denoted by the dashed and solid lines, respectively. Figure 5 shows that the placebo patients' PD severities deteriorate in a much faster rate than the treatment patients as manifested by the departure of two lowess curves, especially at week 40. Figure 5 also reveals one placebo patient (patient 250) who has much worse disease severity than all other patients. This is not surprising because this patient has extremely worse QoL measure as indicated in Figure 1.

To gain further insight into the clinical meanings of the regression parameters β_{10} and β_{11}, we tabulate in Table V the change from baseline to each follow-up visit for disease severity θ_{ij}, the responses QoL, UPDRS, and the odds ratio of the cumulative probability at a certain threshold of the response SEADL. At week 9, the placebo patients are expected to increase 0.162 (95% CI: [0.119, 0.208]) units in disease severity, 1.578 (95% CI: [1.180, 1.980]) units in QoL, 0.999 (95% CI: [0.720, 1.295]) units in UPDRS, and are expected to be 0.811 (95% CI: [0.760, 0.860]) as likely to have cumulative probability at a certain threshold of SEADL, whereas the treatment patients are expected to increase 0.020 (95% CI: [0.001, 0.040]) units in disease severity, 0.200 (95% CI: [0.010, 0.385]) units in QoL, 0.126 (95% CI: [0.006, 0.245]) units in UPDRS, and are expected to be 0.974 (95% CI: [0.950, 0.999]) as likely to have cumulative probability at a certain threshold of SEADL. At week 40, the placebo patients are expected to increase 0.719 (95% CI: [0.528, 0.924]) units in disease severity, 7.013 (95% CI: [5.244, 8.798]) units in QoL, 4.439 (95% CI: [3.198, 5.753]) units in UPDRS, and are expected to be 0.398 (95% CI: [0.295, 0.511]) as likely to have cumulative probability at a certain threshold of SEADL, whereas the treatment patients are expected to increase 0.091 (95% CI: [0.004, 0.177]) units in disease severity, 0.887 (95% CI: [0.043, 1.713]) units in QoL, 0.561 (95% CI: [0.027, 1.087]) units in UPDRS, and are expected to be

Statistics in Medicine

Table V. Change from baseline to each follow-up visit for disease severity θ_{ij}, responses QoL, unified Parkinson's disease rating scale (UPDRS), and the odds ratio (OR) of the cumulative probability at a certain threshold of Schwab and England activities of daily living (SEADL). The number in the subscript is the standard deviation (SD). The numbers within the square brackets are 95% equal-tailed CI.

	θ_{ij}		QoL		UPDRS		OR$\{p(\text{SEADL} \leqslant l)\}$	
	Placebo	Treatment	Placebo	Treatment	Placebo	Treatment	Placebo	Treatment
Week 9	$0.162_{0.022}$ [0.119, 0.208]	$0.020_{0.010}$ [0.001, 0.040]	$1.578_{0.201}$ [1.180, 1.980]	$0.200_{0.097}$ [0.010, 0.385]	$0.999_{0.146}$ [0.720, 1.295]	$0.126_{0.062}$ [0.006, 0.245]	$0.811_{0.025}$ [0.760, 0.860]	$0.974_{0.013}$ [0.950, 0.999]
Week 24	$0.432_{0.060}$ [0.317, 0.554]	$0.055_{0.026}$ [0.003, 0.106]	$4.208_{0.537}$ [3.146, 5.279]	$0.532_{0.258}$ [0.026, 1.028]	$2.664_{0.390}$ [1.919, 3.452]	$0.337_{0.164}$ [0.016, 0.652]	$0.574_{0.048}$ [0.481, 0.668]	$0.932_{0.032}$ [0.871, 0.997]
Week 40	$0.719_{0.099}$ [0.528, 0.924]	$0.091_{0.044}$ [0.004, 0.177]	$7.013_{0.895}$ [5.244, 8.798]	$0.887_{0.429}$ [0.043, 1.713]	$4.439_{0.650}$ [3.198, 5.753]	$0.561_{0.274}$ [0.027, 1.087]	$0.398_{0.055}$ [0.295, 0.511]	$0.890_{0.051}$ [0.795, 0.994]

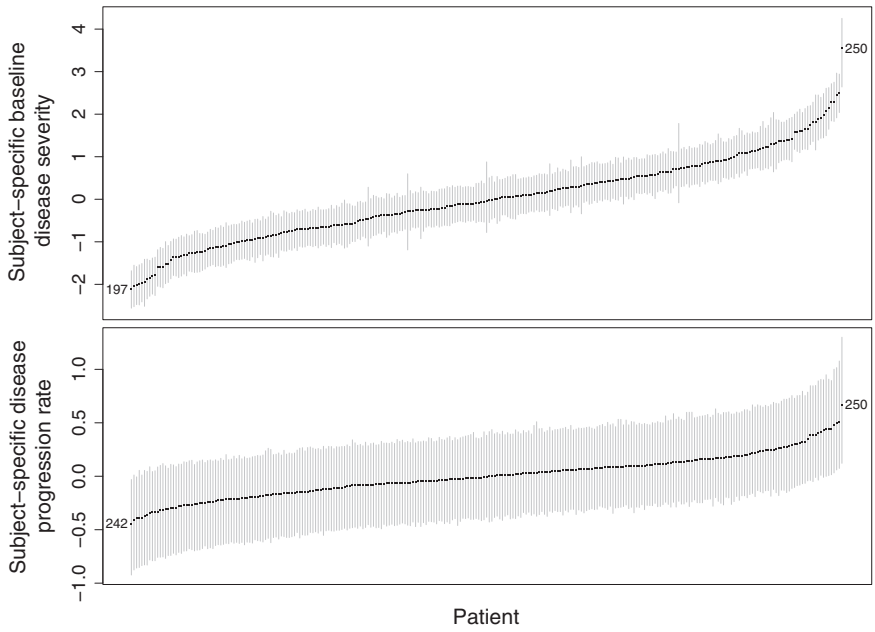

Figure 6. The ranking of subject-specific baseline disease severity (upper panel) and disease progression rate (lower panel) with point estimates and 95% CI. The numbers in the figures are patient numbers.

0.890 (95% CI: [0.795, 0.994]) as likely to have cumulative probability at a certain threshold of SEADL. We omit the results of week 24 and the response fatigue because of space limit, but similar inferences can be made.

In Table IV, we present the SE (σ_u) of the random intercept u_{i1}. The estimate of σ_u from model M_2 (mean: 0.273 and 95% CI: [0.172, 0.419]) indicates the existence of subject-specific disease progression rates. The estimate of the correlation coefficient ρ between the random intercept u_{i0} and the random slope u_{i1} is 0.653 (95% CI: $[0.292, 0.864]$). This suggests that patients whose baseline level of disease severity is worse than that of the average population have a disease progression rate faster than the average and vice versa. To gain further insight into u_{i0}, u_{i1}, and ρ, we plot in Figure 6 the rankings of patients' subject-specific baseline disease severity u_{i0} (upper panel) and disease progression rate u_{i1} (lower panel). Each patient is ordered by his or her rank: patients at the bottom left corner show milder baseline disease severity (higher rank) and slower disease progression rate (higher rank), whereas patients at the upper right corner have poorer baseline disease severity (lower rank) and faster disease progression rate (lower rank). To visualize the effect of high correlation coefficient ρ, we have selected three patients as examples. Patient 250 has the worst baseline disease severity and the fastest disease progression rate. In contrast, patient 197 has the mildest baseline disease severity and he/she ranks No. 6 in disease progression rate. Patient 242 has the slowest disease progression rate while he/she ranks No. 4 in baseline disease severity.

6. Discussion

In this article, we propose a robust model for MLIRTs, in which the robustness against potential outliers in the continuous measurements is achieved by replacing the normal random error distributions by the heavy-tailed normal/independent family of distributions in the common factor submodel. Our simulation studies show that the proposed NI-MLIRT models improve the accuracy of the response-specific parameter estimates and the efficiency of other parameters when outliers exist. On the other hand, the NI-MLIRT models provide comparable results to the MLIRT model when no outliers exist. We apply the proposed method to the motivating Parkinson's disease clinical trial ELLDOPA study and illustrate how the normal/independent distributions can be used to evaluate normality assumption to identify outliers and to obtain more robust inference. We provide subject-specific disease severity estimates for all patients at each visit and the figure to visualize the different disease progression rates in the placebo and treatment groups. We give additional insight into the subject-specific baseline disease

severity and disease progression rate and their correlation. The proposed models can be fitted using standard available software packages such as `R` and `WinBUGS` and can be easily accessible to, modified and extended by practitioners. Please refer to the web-based supporting information[‡] for the code written in `BUGS` language.

Our modeling framework provides great modeling flexibility. For example, a majority of contemporary clinical trials are based on multiple centers or clinics in different geographical locations. The patients recruited by the same center are expected to be correlated, because they are likely to share some common factors, for example, environmental factors. This within-center correlation can be accounted for by adding some center-specific random effects into model (4). Moreover, model (4) assumes linear trend in time. This assumption can be relaxed by adding higher-order terms of time to model the nonlinear disease progression rate.

Our proposed model has some limitations that we view as future research directions. We have assumed that there is a single (unidimensional) latent variable θ_{ij} so that all outcomes measure the underlying disease severity. However, there may be multiple latent variables representing multidimensional (e.g. sensoria, functions, and cognition) impairment caused by PD. Expanding the unidimensional MLIRT model to the multidimensional one is an interesting direction for future research. Note that the discrimination parameter b_k in the MLIRT model controls both within-subject correlation in different outcomes and outcome-specific treatment effect $\boldsymbol{\beta}$ expressed in model (4). If there is low within-subject correlation but a large treatment effect, this model may underestimate the treatment effect and overestimate the correlation [63]. Furthermore, the proposed model does not consider skewness in the continuous responses. However, features of non-normality might be attributed to both skewness and heavy tails. Methods to combine these within a unified framework for the MLIRT models are currently under investigation. Relaxing the normality assumption for the random effects in the MLIRT models using the NI distributions is also part of our future research.

Acknowledgements

This research was supported by two National Institute of Health/National Institute of Neurological Disorders and Stroke grants U01NS043127 and U01NS43128. The authors are grateful to Dr. Barbara C. Tilley, Jordan J. Elm, Adriana Perez, and Ms. Bo He for helpful discussions and comments. Junsheng Ma is supported by the NIH grant 2T32GM074902-06 and by the Lorne C. Bain Endowment.

References

1. Fahn S, Oakes D, Shoulson I, Kieburtz K, Rudolph A, Lang A, Olanow C, Tanner C, Marek K, Parkinson Study Group. Levodopa and the progression of Parkinson's disease. *The New England Journal of Medicine* 2004; **351**(24):2498.
2. Guimaraes P, Kieburtz K, Goetz C, Elm J, Palesch Y, Huang P, Ravina B, Tanner C, Tilley B. Non-linearity of Parkinson's disease progression: implications for sample size calculations in clinical trials. *Clinical Trials* 2005; **2**(6):509–518.
3. O'Brien L, Fitzmaurice G. Analysis of longitudinal multiple-source binary data using generalized estimating equations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2004; **53**(1):177–193.
4. Adams R, Wilson M, Wu M. Multilevel item response models: an approach to errors in variables regression. *Journal of Educational and Behavioral Statistics* 1997; **22**(1):47–76.
5. Anderson T. *An Introduction to Multivariate Statistical Analysis*, 3rd edn., John Wiley & Sons: Hoboken, New Jersey, 2003.
6. Andersen E. Latent regression analysis based on the rating scale model. *Psychology Science* 2004; **46**:209–226.
7. Christensen K, Bjorner J, Kreiner S, Petersen J. Latent regression in loglinear Rasch models. *Communications in Statistics-Theory and Methods* 2004; **33**(6):1295–1313.
8. Mislevy R. Estimation of latent group effects. *Journal of the American Statistical Association* 1985; **80**:993–997.
9. Zwinderman A. A generalized Rasch model for manifest predictors. *Psychometrika* 1991; **56**(4):589–600.
10. Maier K. A Rasch hierarchical measurement model. *Journal of Educational and Behavioral Statistics* 2001; **26**(3):307–330.
11. Kamata A. Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement* 2001; **38**(1):79–93.
12. Fox J. Multilevel IRT using dichotomous and polytomous response data. *British Journal of Mathematical and Statistical Psychology* 2005; **58**(1):145–172.
13. Rijmen F, Tuerlinckx F, De Boeck P, Kuppens P. A nonlinear mixed model framework for item response theory. *Psychological Methods* 2003; **8**(2):185.
14. Fox J, Glas C. Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika* 2001; **66**(2):271–288.
15. Fox J. Applications of multilevel IRT modeling. *School Effectiveness and School Improvement* 2004; **15**(3-4):261–280.

[‡]*Supporting information may be found in the online version of this article.*

16. Fox J. Multilevel IRT modeling in practice with the package mlirt. *Journal of Statistical Software* 2007; **20**(5):1–16.
17. Fox J. *Bayesian Item Response Modeling: Theory and Applications*. Springer Verlag: New York, New York, 2010.
18. Natesan P, Limbers C, Varni J. Bayesian estimation of graded response multilevel models using Gibbs sampling: Formulation and illustration. *Educational and Psychological Measurement* 2010; **70**(3):420–439.
19. Hung L, Wang W. The generalized multilevel facets model for longitudinal data. *Journal of Educational and Behavioral Statistics* 2012; **37**:231–255.
20. Van der Linden W, Hambleton R. *Handbook of Modern Item Response Theory*. Springer Verlag: New York, New York, 1997.
21. Lord F, Novick M, Birnbaum A. *Statistical Theories of Mental Test Scores*. Addison-Wesley: Boston, MA, 1968.
22. Pinheiro J, Liu C, Wu Y. Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics* 2001; **10**(2):249–276.
23. Lachos V, Bandyopadhyay D, Dey D. Linear and nonlinear mixed-effects models for censored HIV viral loads using normal/independent distributions. *Biometrics* 2011; **67**:1594–1604.
24. Mislevy R, Bock R. Biweight estimates of latent ability. *Educational and Psychological Measurement* 1982; **42**(3):725–737.
25. Schuster C, Yuan K. Robust estimation of latent ability in item response models. *Journal of Educational and Behavioral Statistics* 2011; **36**(6):720–735.
26. Pison G, Rousseeuw P, Filzmoser P, Croux C. Robust factor analysis. *Journal of Multivariate Analysis* 2003; **84**(1):145–172.
27. Salibian-Barrera M, Van Aelst S, Willems G. Principal components analysis based on multivariate MM estimators with fast and robust bootstrap. *Journal of the American Statistical Association* 2006; **101**(475):1198–1211.
28. Hubert M, Van Driessen K. Fast and robust discriminant analysis. *Computational Statistics & Data Analysis* 2004; **45**(2):301–320.
29. Liu C. Bayesian robust multivariate linear regression with incomplete data. *Journal of the American Statistical Association* 1996; **91**:1219–1227.
30. Lange K, Sinsheimer J. Normal/independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics* 1993; **2**:175–198.
31. Sutradhar B, Ali M. Estimation of the parameters of a regression model with a multivariate t error variable. *Communications in Statistics-Theory and Methods* 1986; **15**(2):429–450.
32. Lange K, Little R, Taylor J. Robust statistical modeling using the t distribution. *Journal of the American Statistical Association* 1989:881–896.
33. Lachos V, Bandyopadhyay D, Garay A. Heteroscedastic nonlinear regression models based on scale mixtures of skew-normal distributions. *Statistics and Probability Letters* 2011; **81**:1208–1217.
34. Rosa G, Padovani C, Gianola D. Robust linear mixed models with normal/independent distributions and Bayesian MCMC implementation. *Biometrical Journal* 2003; **45**(5):573–590.
35. Lin T, Lee J. A robust approach to t linear mixed models applied to multiple sclerosis data. *Statistics in Medicine* 2006; **25**(8):1397–1412.
36. Lin T, Lee J. Bayesian analysis of hierarchical linear mixed modeling using the multivariate t distribution. *Journal of Statistical Planning and Inference* 2007; **137**(2):484–495.
37. Ho H, Lin T. Robust linear mixed models using the skew t distribution with application to schizophrenia data. *Biometrical Journal* 2010; **52**(4):449–469.
38. Russo C, Paula G, Aoki R. Influence diagnostics in nonlinear mixed-effects elliptical models. *Computational Statistics and Data Analysis* 2009; **53**(12):4143–4156.
39. Meza C, Osorio F, De la Cruz R. Estimation in nonlinear mixed-effects models using heavy-tailed distributions. *Statistics and Computing* 2012; **22**:121–139.
40. Li N, Elashoff R, Li G. Robust joint modeling of longitudinal measurements and competing risks failure time data. *Biometrical Journal* 2009; **51**(1):19–30.
41. Huang X, Li G, Elashoff R. A joint model of longitudinal and competing risks survival data with heterogeneous random effects and outlying longitudinal measurements. *Statistics and Its Interface* 2010; **3**(2):185.
42. Lee S, Xia Y. A robust Bayesian approach for structural equation models with missing data. *Psychometrika* 2008; **73**(3):343–364.
43. Abanto-Valle C, Bandyopadhyay D, Lachos V, Enriquez I. Robust Bayesian analysis of heavy-tailed stochastic volatility models using scale mixtures of normal distributions. *Computational statistics and data analysis* 2010; **54**(12): 2883–2898.
44. Abanto-Valle C, Migon H, Lachos V. Stochastic volatility in mean models with scale mixtures of normal distributions and correlated errors: a Bayesian approach. *Journal of Statistical Planning and Inference* 2011; **141**(5):1875–1887.
45. Osorio F, Paula G, Galea M. On estimation and influence diagnostics for the Grubbs' model under heavy-tailed distributions. *Computational Statistics and Data Analysis* 2009; **53**(4):1249–1263.
46. Ghosh P, Bayes C, Lachos V. A robust Bayesian approach to null intercept measurement error model with application to dental data. *Computational Statistics and Data Analysis* 2009; **53**(4):1066–1079.
47. Lachos V, Angolini T, Abanto-Valle C. On estimation and local influence analysis for measurement errors models under heavy-tailed distributions. *Statistical Papers* 2011; **52**(3):567–590.
48. Cao C, Lin J, Zhu X. On estimation of a heteroscedastic measurement error model under heavy-tailed distributions. *Computational Statistics and Data Analysis* 2012; **56**:438–448.
49. Schrag A. Quality of life and depression in Parkinson's disease. *Journal of the Neurological Sciences* 2006; **248**(1):151–157.
50. Schifitto G, Friedman J, Oakes D, Shulman L, Comella C, Marek K, Fahn S, The Parkinson Study Group ELLDOPA. Fatigue in levodopa-naive subjects with Parkinson disease. *Neurology* 2008; **71**(7):481–485.

51. Schwab R, England A. Projection technique for evaluating surgery in Parkinson's disease. In *Third Symposium on Parkinson's Disease*. Livingstone: Edinburgh, 1969; 152–157.
52. Lord F. *Applications of Item Response Theory to Practical Testing Problems*. L. Erlbaum Associates: Hillsdale, NJ, 1980.
53. Samejima F. *Graded Response Model*, chap. 5. Springer: New York, New York, 1997; 85–100.
54. Curtis S. BUGS code for item response theory. *Journal of Statistical Software* August 2010; **36**:1–34.
55. Andrews D, Mallows C. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)* 1974; **36**:99–102.
56. Gelman A, Carlin J, Stern H, Rubin D. *Bayesian Data Analysis*. CRC press: Boca Raton, FL, 2004.
57. Geisser S. *Predictive Inference: An Introduction*, Vol. 55. Chapman & Hall/CRC: Boca Raton, FL, 1993.
58. Dey D, Chen M, Chang H. Bayesian approach for nonlinear random effects models. *Biometrics* 1997; **53**:1239–1252.
59. Sinha D, Dey D. Semiparametric Bayesian analysis of survival data. *Journal of the American Statistical Association* 1997; **92**:1195–1212.
60. Carlin B, Louis T. *Bayesian Methods for Data Analysis*. Chapman & Hall/CRC: Boca Raton, FL, 2009.
61. Spiegelhalter D, Best N, Carlin B, Van der Linde A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B:Statistical Methodology* 2002; **64**(4):583–639. DOI: 10.1111/1467-9868.00353.
62. Cleveland W. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association* 1979; **74**:829–836.
63. Dunson D. Bayesian methods for latent trait modelling of longitudinal data. *Statistical Methods in Medical Research* 2007; **16**(5):399.