Published in final edited form as:
Comput Stat Data Anal. 2007 August 15; 51(12): 5718–5730. doi:10.1016/j.csda.2006.09.036.

Efficient Hybrid EM for Linear and Nonlinear Mixed Effects Models with Censored Response

Florin Vaida[*], **Anthony P. Fitzgerald**[†], and **Victor DeGruttola**[‡]

[*]Department of Family and Preventive Medicine, UC San Diego School of Medicine, La Jolla, CA 92093-0717, USA; email: vaida@ucsd.edu [†]Department of Epidemiology and Public Health, Brookfield Health Sciences Complex, University College Cork, College Road, Cork, Ireland
[‡]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA

Abstract

Medical laboratory data are often censored, due to limitations of the measuring technology. For pharmacokinetics measurements and dilution-based assays, for example, there is a lower quantification limit, which depends on the type of assay used. The concentration of HIV particles in the plasma is subject to both lower and upper quantification limit. Linear and nonlinear mixed effects models, which are often used in these types of medical applications, need to be able to deal with such data issues. In this paper we discuss a hybrid Monte Carlo and numerical integration EM algorithm for computing the maximum likelihood estimates for linear and non-linear mixed models with censored data. Our implementation uses an efficient block-sampling scheme, automated monitoring of convergence, and dimension reduction based on the QR decomposition. For clusters with up to two censored observations numerical integration is used instead of Monte Carlo simulation. These improvements lead to a several-fold reduction in computation time. We illustrate the algorithm using data from an HIV/AIDS trial. The Monte Carlo EM is evaluated and compared with existing methods via a simulation study.

Keywords

Monte Carlo EM; HIV-1 viral dynamics; quantification limit; LME; NLME; likelihood estimation

1 Introduction

When analyzing medical data, the statistician is often confronted with censored observations. For laboratory data, these may be due to limitations of the measuring technology. In pharmacokinetics the concentration of drug in plasma is subject to a limit of quantification below which the measurement is not reliable, or even possible. Similarly, the HIV-1 viral load, which is currently the primary marker of HIV infection, has a lower and a upper quantification limit, which depend on the type of assay used. The viral load of patients receiving anti-retroviral treatment will typically decline and stay for a longer period of time below the lower limit of quantification.

Since in biomedical applications the observations are often non-linear and longitudinal, non-linear mixed effects models (NLME) are a popular modelling tool for these data. In practice, the censoring problem is ignored, or dealt with in an ad-hoc way. In this paper we use some novel EM computational techniques in order to adjust for censored responses in NLME estimation. The E-step uses numerical integration, for clusters with up to two censored observations, or Monte Carlo integration for clusters with more than two censored observations. As such our algorithm is a hybrid between Monte Carlo EM (MCEM) and a "classical" EM using numeric integration. We call it a hybrid EM (HEM). In this algorithm the data augmentation scheme involves both the random effects and the censored observations. An alternative computational method is multiple imputation (Rubin, 1996; Fitzgerald *et al.*, 2002, MI). While HEM has the potential of more precise estimation of the MLE for censored data, MI enjoys straightforward implementation using existing NLME software, such as the nlme suite for R/S-plus (Pinheiro and Bates, 2000), or PROC NLMIXED in SAS (Wolfinger, 1999). We compare the two methods and show here that both methods are superior to ad-hoc approaches, such as using the censoring limits as observed values. The end user has the ultimate choice in the trade-off between precision and ease of implementation. This choice has been heavily influenced by the absence of ready-to-use software for NLME with censored response, although Hughes (1999) also made available software for linear mixed-effects models (LME) with censored response. For those choosing HEM we provide a versatile, self-monitoring and computationally efficient program implemented in R. The improvements with respect to the "state of the art" include: automatic monitoring of convergence based on an approximate likelihood objective function; automatic choice of Monte Carlo sample size; block-sampling of the censored data and random effects; efficient computation using dimension reduction using QR decomposition; incorporating the linearization step in the EM loop. In addition, we applied the same improvements to an algorithm for LME with censored response.

We illustrate the general methodology developed here to the analysis of an AIDS clinical trial. In ACTG 315 study (Lederman *et al.*, 1998) the viral dynamics are nonlinear, and later viral load observations are often below the limit of quantification of the assay (left-censored). A second situation (analysis not presented) regards modelling the setpoint HIV-1 RNA levels of untreated individuals with acute HIV infection from the Acute Infection and Early Disease Research (AIEDRP) study. Here observations taken in the acute stage of infection are often *above* the limit of quantification of the assay (right-censored).

The likelihood of NLME models with completely observed response is untractable, and the MLE is not available in closed form. Briefly stated, NLME are solved by iteratively linearizing the mean function using a Taylor expansion, followed by a linear-mixed-effects step (Laird and Ware, 1982; Lindstrom and Bates, 1990). Several linearization methods have been proposed: Sheiner and Beal (1980); Lindstrom and Bates (1990); Wolfinger (1993); Kiuchi *et al.* (1995); Pinheiro and Bates (1995). In each case the resulting solution is an approximate MLE. Pinheiro and Bates (1995) concluded based on a comparative study that the method of Lindstrom and Bates (1990) using iterative linearization around the current estimates for the parameter and random effects estimates performs well. For a detailed account of the NLME see the recent monographes of Davidian and Giltinan (1995), Vonesh and Chinchilli (1997), and Pinheiro and Bates (2000). The issue of censored response for a LME was considered by Hughes (1999), who used a Monte Carlo EM algorithm extending the methods of Laird and Ware (1982). For NLME our work builds on Fitzgerald (2000). Wu (2002, 2004) has extended the work of Hughes (1999) to LME and NLME which also accommodate error in variables. Beal (2001) discusses practical issues related to left-censored observations in pharmacokinetics and compares several methods for dealing with them in fixed-effects modeling.

2 Monte Carlo EM for Linear Mixed Effects models with Censored Response

After briefly summarizing Hughes' Monte Carlo EM algorithm for LME, we describe our computationally efficient implementation, including a simple and general framework for automatic selection of Monte Carlo sample size and monitoring convergence of the HEM. This forms the basis for the algorithm for NLME with censored response, presented in the next section.

2.1 Hughes' algorithm

Hughes (1999) proposed a MCEM algorithm for LME with censored data. Consider the Laird-Ware linear mixed-effects model

$$y_i = X_i\beta + Z_i b_i + e_i, \tag{1}$$

$i = 1, \ldots, m$ with b_i and $e_i = (e_{i1}, \ldots, e_{in_i})^\top$ given by

$$b_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2 D\right), \qquad e_{ij} \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right), \tag{2}$$

independent of each other. D is a positive definite matrix depending on a vector of parameters γ. Write $\sigma^2 D = \Psi$ and note that $V_i = \text{Var}(y_i) = Z_i \Psi Z_i^\top + \sigma^2 I$. In the settings which interest us here the response y_{ij} is not fully observed for all i, j. Let the observed data for the i^{th} subject be (Q_i, C_i), where Q_i represents the vector of uncensored readings or censoring level, and C_i the vector of censoring indicators:

$$\begin{aligned} y_{ij} &\le Q_{ij} \quad \text{if} \quad C_{ij} = 1 \\ y_{ij} &= Q_{ij} \quad \text{if} \quad C_{ij} = 0. \end{aligned} \tag{3}$$

We will assume for simplicity of description that the data are left-censored. The extensions to arbitrary (left, right, or interval) censoring are immediate.

Hughes (1999) modified the Laird and Ware (1982) EM equations to incorporate censoring. At the M-step, these equations are: $\widehat{\beta} = \sum_{i=1}^m \left(X_i^\top V_i^{-1} X_i\right)^{-1} \left\{\sum_{i=1}^m X_i^\top V_i^{-1} E(y_i | C_i, Q_i)\right\}$; $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^m E\left(\widehat{e}_i^\top \widehat{e}_i | C_i, Q_i\right)$; $\widehat{\Psi} = \frac{1}{m}\sum_{i=1}^m E\left(\widehat{b}_i \widehat{b}_i^\top | C_i, Q_i\right)$, where $\hat{e}_i = y_i - X_i\hat{\beta} - Z_i\hat{b}_i$, $n = \sum_{i=1}^m n_i$, and

$$\widehat{b}_i = \left(D^{-1} + Z_i^\top Z_i\right)^{-1} Z_i^\top (y_i - X_i\beta). \tag{4}$$

The expectations are taken at the current parameter value θ; $\hat{\beta}$ is updated using as missing data $\{y_{ij} : C_{ij} = 1\}$, but not b_i; whereas $\hat{\sigma}^2$, $\hat{\Psi}$ are updated with $\{y_{ij} : C_{ij} = 1\}$ and b_i as missing data. Strictly speaking, this is not an EM but rather a SAGE algorithm (Meng and van Dyk, 1997).

The conditional expectations in Hughes' equations are functions of $E(y_i | Q_i, C_i, \theta)$ and $\text{Var}(y_i | C_i, Q_i, \theta)$. These are computed at the E-step by simulating y_i from the marginal distribution of

y_i, $p(y_i|C_i, Q_i, \theta)$, which is truncated multivariate normal, using a Gibbs sampler. Upon convergence b_i is estimated by the empirical Bayes estimator $E(\hat{b_i}|Q_i, C_i)$. The variance of the MLE $\hat{\theta}$, estimated at convergence, is adjusted for the censored information using Louis' formula (Orchard and Woodbury, 1972; Louis, 1982). The variance of the fixed effects in the approximate MLE is given (Hughes, 1999) by

$$\mathrm{Var}(\widehat{\beta})=\left(\sum_{i=1}^{m}\left\{X_i^\top V_i^{-1}X_i - X_i^\top V_i^{-1}\mathrm{Var}(y_i|Q_i,C_i)V_i^{-1}X_i\right\}\right)^{-1}. \quad (5)$$

2.2 The proposed Hybrid EM: E-step

The EM implementation we propose for the LME with censored data differs from Hughes' in several respects, in both E- and M-steps.

At the E-step, we treat differently clusters i with 0, 1, or two censored observations from those with 3 or more censored observations. In the first case, the conditional mean and variance of censored data are calculated in closed form, without the use of Gibbs sampling, using formulae for bivariate truncated normal (Maddala, 1996) and the mvtnorm package in R (Genz, 1992). These are then used in the M-step formulas, discussed below.

For clusters with 3 or more censored observations we use Monte Carlo simulation. Instead of sampling y_i from its marginal distribution as Hughes (1999), we sample (y_i, b_i) using a block Gibbs sampler, as follows:

1. Sample $y_i \sim p(y_i|b_i, C_i, Q_i, \theta)$. Conditional on b_i, y_i is a vector of independent observations, whose distributions are truncated normal, each with untruncated variance σ^2 and untruncated mean $x_{ij}^\top\beta+z_{ij}^\top b_i$, on the interval $\{y_{ij} \le Q_{ij}\}$ (x_{ij} and z_{ij} are the jth rows of X_i and Z_i, respectively).
2. Sample $b_i \sim p(b_i|y_i, Q_i, C_i, \theta) = p(b_i|y_i, \theta)$. The target distribution is multivariate normal with mean $\hat{b_i}$ as in (4) and variance $\mathrm{Var}(b_i|y_i)=\sigma^2(D^{-1}+Z_i^\top Z_i)^{-1}$. Note that the entire vector y_i is used for simulating b_i, not only the censored components.

While this additional data augmentation in the Gibbs sampler has potentially slower mixing for the sampled y_i's, it greatly simplifies computations. Hughes' Gibbs sampler requires an update of the mean and variance of the fully conditional distribution of each sampled y_{ij}. This involves costly matrix multiplication and inversion. More importantly, there are advantages in sampling y_i in block: the vector sampler in R is ten times faster than the univariate sampler for the independent components of the same vector, as it uses the fast implementations of the R vector functions `pnorm(log.p=T)` and `qnorm(log.p=T)`.

The variance of b_i and the matrices used in computing the mean of b_i are of dimension $q \times q$ (dimension of b_i), and can be computed once at the beginning of the entire Gibbs sampler. They are also used in the M-step of the algorithm, so they induce no additional computational cost.

Using different E-step schemes for different clusters is possible because the clusters are independent. The hybrid EM method improves computation time in two ways: firstly, the computation of the E-step is faster than the MCEM version; secondly, since there is no Monte Carlo error it improves the convergence of the EM and therefore converges in fewer steps than the standard MCEM.

2.3 The proposed hybrid EM: M-step

For the LME with fully-observed data, Bates and Pinheiro (1997) and Schafer (1998) show that σ^2 can be updated at the M-step based on the marginal likelihood, which leads to faster convergence of the EM algorithm, following the observation of Lindstrom and Bates (1990). Using this idea in the context of censored data, we update β, σ^2 with $\{y_{ij} : C_{ij} = 1\}$ as missing data, and Ψ using $\{y_{ij} : C_{ij} = 1\}$ and b_i as missing data. Using the "pseudo-data" notation (see, e.g. Pinheiro and Bates, 2000, p.63) we decompose $D^{-1} = \Delta^\top \Delta$ and write:

$$\delta=(\beta^\top, b_1^\top, \ldots, b_m^\top)^\top, \widetilde{y}=(\widetilde{y}_1^\top, \ldots, \widetilde{y}_m^\top)^\top,$$

$$\begin{pmatrix} \widetilde{y}_i & \widetilde{X}_i & \widetilde{Z}_i \end{pmatrix} = \begin{pmatrix} y_i & X_i & Z_i \\ 0 & 0 & \Delta \end{pmatrix}, \quad \text{and} \quad M=\begin{pmatrix} \widetilde{X}_1 & \widetilde{Z}_1 & & \\ \vdots & & \ddots & \\ \widetilde{X}_m & & & \widetilde{Z}_m \end{pmatrix}. \tag{6}$$

Then, the M-step updates are:

$$\widehat{\delta}=(M^\top M)^{-1} M^\top E(\widetilde{y}) \tag{7}$$

$$\widehat{\sigma}^2=\frac{1}{n}\|E(\widetilde{y}) - M\widehat{\delta}\|^2+\frac{1}{n}\sum_{i=1}^{m}\mathrm{tr}\{\mathrm{Var}(y_i)\} - \frac{1}{n}\sum_{i=1}^{m}\mathrm{tr}\left\{W_i Z_i^\top \mathrm{Var}(y_i) Z_i\right\} \tag{8}$$

$$\widehat{\Psi}=\frac{1}{m}\sum_{i=1}^{m}E(b_i b_i^\top)=\frac{1}{m}\sum_{i=1}^{m}E(b_i)E(b_i)^\top+\frac{1}{m}\mathrm{Var}(b_i) \tag{9}$$

where $W_i=(Z_i^\top Z_i+D^{-1})^{-1}$, $E(b_i)=W_i Z_i^\top \{E(y_i) - X_i\beta\}$, $\mathrm{Var}(b_i)=\sigma^2 W_i+W_i Z_i^\top \mathrm{Var}(y_i) Z_i W_i$, and $E(y_i)$, $\mathrm{Var}(y_i)$ are the mean and variance conditional on $\{C_i, Q_i ; i = 1 \ldots m\}$, taken at the current parameter value $\theta = (\beta, \sigma^2, D)$. The computations use dimension reduction based on QR decomposition, as described in Bates and Pinheiro (1997) and Pinheiro and Bates (2000).

In (9) we assumed that Ψ is unstructured. When Ψ is assumed diagonal, the updated Ψ is a diagonal matrix with same diagonal elements as the right hand side in (9).

2.4 Automatic monitoring of convergence

For clusters with three or more censored observations the proposed HEM behaves like a MCEM. Choosing the Monte Carlo (MC) sample size and monitoring MCEM convergence are important issues which haven't received yet a satisfactory resolution. Chan and Ledolter (1995) show that the likelihood sequence of the EM stabilizes to an approximate AR(1) process, with variance inversely proportional to the MC sample size G. Booth and Hobert (1999) suggest choosing G by comparing the MC error with the asymptotic error of the MLE. This requires expensive computations of variance matrices, or second order derivatives, at each step of the algorithm. Vaida and Meng (2005) use a gradual increase of G to ensure a "smooth" transition of the MCEM to the new stationary distribution for large G. They take the final estimate of the MLE by averaging the samples run in the "plateau" stage of the MCEM. Here we propose a simple general approach to MCEM convergence.

The objective function for monitoring EM convergence is the log-likelihood, which is monotonously increasing to its MLE value. However, in many applications this is not easy to compute. Alternatively, convergence can be decided when the change in parameters is "small enough". It is not clear what metric to use to combine the parameters when determining the change; the log-likelihood is a natural choice. We are interested here in an objective function that is one-dimensional, easy to compute, and which does not add to the computational burden. For LME with censored data we use the form of the log-likelihood for the corresponding LME, computed as a function of the parameters alone (e.g. Pinheiro and Bates, 2000, formula (2.13)). In our case the sequence of values of the objective function (called by abuse of language "log-likelihood") still follows an approximate AR(1) process upon convergence.

The MLE solution given by MCEM is approximate, up to the MC error. In our experience, the MC error of the log-likelihood is dominating by far the error due to non-convergence of the EM. Instead of engaging in a losing race of making MCEM look like a "bona fide" EM by using extremely large values of G, we decided to embrace this "character flaw" of MCEM and work with, rather than against, its variability. We propose to declare convergence when the empirical standard deviation of the MCEM objective function, at stationarity, has a predetermined, small value, s^*. The value of G can be adjusted "on the fly" to achieve this goal. Thus, the algorithm has several stages, as follows.

In the *burn-in stage* the parameter converges to the vicinity of the MLE. For this we use a small MCMC sample size G_1, e.g., $G_1 = 100$. Call l_i the log-likelihood at step i. We end the burn-in stage when $l_{i+1} < l_i$, i.e. when the log-likelihood starts to jump up-and-down. This usually takes a small number of steps. In the second stage we evaluate the standard error of l_i. Keeping $G = G_1$ we run the algorithm for I_2 steps, e.g., $I_2 = 10$, and compute the standard deviation of the log-likelihood in this stage, using the AR(1) assumption. We compare this standard deviation, s_{l2}, with the tolerance of MC standard deviation, s^*. The approximate MC sample size needed to achieve s^* is $G^* = (s_{l2}/s^*)^2 G_1$. Next, we start the *transition stage* between $G = G_1$ and $G = G^*$. The increase in G needs to be smooth so that the EM doesn't spend too much computation time far away from the MLE. We increase G such that the decrease in s_l is linear. After each increase G is kept fixed until l_i shows a change in direction, then G is increased again, and so forth, until reaching G^*. This ensures running the algorithm in a state of "quasi-stationarity". Finally, in the *plateau stage* run the algorithm for a number of steps I_4, e.g., $I_4 = 10$, at $G = G^*$. At this point we can either stop the algorithm, or evaluate s_{l4} based on the last I_4 steps and if $s_{l4} > s_l^*$, again increase G accordingly and run it for I_5 steps. The parameter estimates are those from the last step. A multiple stage MCEM including burn-in, transition and plateau stages was discussed by Vaida and Meng (2005), but without automatic monitoring. The notion of burn-in is borrowed from the MCMC literature (see, e.g., Gilks *et al.*, 1996).

This algorithm balances the monotonicity of the EM with the variability of MCMC. During the transition period we assume that the incremental difference due to EM is larger than the MCMC error as long as the log-likelihood sequence is monotone. When the sequence is no longer monotone the MCMC sample size can be increased. We recommend this as a general strategy for monitoring MCEM convergence.

3 Nonlinear mixed effects models with censored response

We apply now the ideas and algorithm designed for LME to NLME with censored response. Extending the notation of previous section, consider the general NLME model:

$$y_{ij} = f(\beta, b_i) + e_{ij}. \tag{10}$$

The conditional mean of y_{ij}, $f(\beta, b_i) = f(\beta, b_i, x_{ij})$ is a non-linear function of the fixed parameter β and of the random effect vector b_i; x_{ij} is a vector of covariates, and b_i and e_{ij} are given by (2). For example, $f(\beta, b_i)$ may be given by (14), with $b_i = \beta_i - \beta$, and $x_{ij} = t_{ij}$. The marginal likelihood for the NLME, $\mathrm{Lik}(\beta,\sigma^2,\gamma)=\prod_{i=1}^{m}\int p(y_i\,|b_i,\beta,\sigma^2)p(b_i|\,\sigma^2 D)db_i$ is in general not in closed form. Most algorithms for computing the MLE $(\hat{\beta}, \hat{\sigma^2}, \hat{\gamma})$ and empirical Bayes estimators (predictors) for the random effects $\hat{b_i}$, rely on iteratively linearizing the conditional mean function and solving the resulting LME model. Our algorithm for NLME with censored response deal with the censored data within the LME step.

For NLME with complete response, if the current estimates for (β, b_i) are (β^*, b_i^*), the linearization step yields the LME

$$w_i = X_i^*\beta + Z_i^* b_i + e_i, \tag{11}$$

$i = 1, \ldots, m$, where

$$w_i = y_i - \{f_i^* - \left(\frac{\partial f_i^*}{\partial \beta}\right)\beta^* - \left(\frac{\partial f_i^*}{\partial b_i}\right) b_i^*\}, \tag{12}$$

$X_i^* = \frac{\partial f_i^*}{\partial \beta}$, $Z_i^* = \frac{\partial f_i^*}{\partial b_i}$, y_i is the n_i-vector dependent variable for the i^{th} subject, f_i, e_i are respectively the corresponding mean function and error n_i-vectors, and the starred terms are computed at (β^*, b_i^*). The MLE for the LME model (11) yields updated estimates of $(\beta, \sigma^2, \gamma)$ and b_i, and the algorithm is iterated to convergence (Wolfinger, 1993). The LME step may be solved using a Newton-Raphson algorithm (Lindstrom and Bates, 1990;Pinheiro and Bates, 2000) or EM-type algorithms (Laird and Ware, 1982).

Lindstrom and Bates (1990) precede the linearization step by a penalized non-linear least squares (PNLS) step: for given variance matrix D, $\hat{\beta}$ and $\hat{b_i}$ are updated by minimizing the objective function

$$\sum_{i=1}^{m}(y_i - f_i(\beta,b_i))^\top (y_i - f_i(\beta,b_i)) + b_i^\top D^{-1} b_i. \tag{13}$$

The PNLS leads to the same solution as the iterative-LME algorithm without the PNLS step, possibly faster in some cases (Bates and Pinheiro, 1997).

Assume now that the response y_{ij} is not fully available, but rather the censoring value and indicator, (Q_{ij}, C_{ij}) are observed, as in (3). We apply the EM algorithm for LME with censored response, but with each EM iteration preceded by a linearization step. More specifically, the algorithm goes as follows:

1. Linearization step. Compute w_i, X_i^*, Z_i^*, based on current parameter estimates θ^*, as in (11). In this calculation use the censoring limits Q_{ij} instead of y_{ij}. This generates a LME with censored response (w_{ij}, C_{ij}).

2. E-step. Same as the simulation E-step from section 2.2. Compute the expectations numerically, or simulate the fully observed response of the linearization step LME and the random effects using a block Gibbs sampler.

3. M-step. Update the parameter values as in section 2.3, using (7)–(9).

The algorithm is iterated to convergence. The starting values are provided by an NLME where the censored data are replaced with ad-hoc values, e.g. the truncation limit, or half the truncation limit (Wu and Ding, 1999). Upon convergence b_i is estimated by $\hat{b_i}$ from (7). The variance of the MLE $\hat{\theta}$ is estimated at convergence, based on the linearized model. Pinheiro and Bates (2000) and Bates and Watts (1988) discuss the accuracy of this approximation for NLME. As for LME, the variance is adjusted for the censored information using Louis' formula. The variance of the fixed effects in the approximate MLE is given (Hughes, 1999) by (5). This is computed upon convergence, using a hybrid approach similar to the E-step of the EM algorithm.

Note that we incorporated the linearization step as part of the EM algorithm. The NLME algorithm in section 3.1 suggests that each linearization step be followed by the computation of the full MLE of the linearized model, which is what Fitzgerald (2000) and Wu (2002) did. The algorithm we present here is equivalent to a single EM iteration for the linearized model. Both versions lead to the same MLE solution, since both the linearization step and the EM step need to converge in order for the whole algorithm to converge. The choice of how many EM steps to use between linearization steps is driven by computational efficiency. In our experience, the sampling E-step is much more time consuming than the linearization step, and thus a re-linearization for each EM iteration is warranted.

The choice of Monte Carlo sample size and monitoring the convergence of the MCEM are detailed in section 2.4. The objective function, following Lindstrom and Bates (1990), is the log-likelihood of the linearized LME, as a function of the parameters, see also equation (2.13) in Pinheiro and Bates (2000). There is, however, one specific difference for the NLME. Because of the linearization step, even if the E-step had no Monte Carlo error at all, the log-likelihood sequence would not necessarily be increasing, rather it may have occasional “turns”, or even converge decreasingly towards the MLE limit. For this reason we need a different mechanism to identify the end of the burn-in stage. Such a mechanism is provided by the MC standard deviation of the log-likelihood process. We compute this standard deviation s_{l1} based on a batch of, say, 10 successive steps, using an AR(1) approximation. If s_{l1} is decreasing over two successive batches this indicates that the process is not stationary yet, and the burn-in is not over. When, due to the random error in the Monte Carlo step, s_{l1} is increasing over two successive values, this indicates stationarity and we can proceed to the transition stage. The target MC sample size is based on the final value of s_{l1} from burn-in.

4 An AIDS Study

In this application we reanalyze the HIV viral load data from clinical trial ACTG 315 Lederman *et al.* (1998). The HIV viral load is the primary measure of HIV infection. Commercially available assays have lower limits of accurate quantification of between 40 and 400 copies/mL of plasma. The observations below this quantification limit (QL) are censored at this value. Wu and Ding (1999) used a non-linear mixed effects models (NLME) in a statistical analysis of HIV-1 viral decay data after initiation of ARV. This model was used in subsequent publications, including Wu (2004) who also considered measurement error in the CD4 covariate process. The bi-exponential model of Wu and Ding (1999) was rooted in the mathematical model for HIV proposed by Perelson *et al.* (1996). The model is

$$y_{ij}=\log_{10}\left\{P_{1i}\exp\left(-\delta_i t_{ij}\right)+P_{2i}\exp\left(-\lambda_i t_{ij}\right)\right\}+e_{ij}, \quad (14)$$

where y_{ij} is the $\log_{10}$ HIV-1 RNA viral load for the i^{th} subject at time t_{ij}. The subject-specific random effects and the error terms satisfy (2) with $\beta_i = \beta + b_i = (\ln P_{1i}, \ln P_{2i}, \delta_i, \lambda_i)^\top$, $\beta = (\ln P_1, \ln P_2, \delta, \lambda)^\top$. The parameters δ and λ are two viral elimination rates of the two components, corresponding in theory to two different pools of HIV infected cells. It is assumed that $\delta \gg \lambda$. The first component dominates the viral dynamics, lasting roughly for the first two weeks of ARV treatment, and the second component controls the second phase, of more shallow viral decay, lasting roughly between weeks 2 and 12 of treatment. The data are from the first 12 weeks of follow-up of HIV clinical trial ACTG 315 (see Lederman *et al.*, 1998, for details). In this paper we are only concerned with modelling the viral decay phase, prior to viral rebound. Observations following viral rebound (a 1 $\log_{10}$ increase over the nadir value) were not included. See Fitzgerald *et al.* (2002) for an analysis including the viral rebound. For the observations below the limit of quantification of the NASBA assay of 100 copies/mL, y_{ij} is left-censored at $\log_{10}$ 100. Since the censoring value is constant (type 1 censoring), the censoring is independent of the complete data. Wu and Ding dealt with the censored data in an ad-hoc manner, by replacing it with half the QL value (HQL). The data set for our analysis is slightly different than the one used by Wu and Ding (1999), who kept only the first censored value for each subject in the analysis. Measurements were taken at days 0, 2, 7, 10, and weeks 2, 3, 4, 8, 12. (Figure 1). The data consist of 381 observations on 47 subjects. Nineteen subjects had at least one censored observation (11 had one, 4 had two and 4 had 3 or more censored observations).

The results of this analysis are presented in Table 1. The coefficient estimates are similar for HEM and HQL. The turnover rate for the second component, λ, whose estimation is based mostly on the later follow-up observations is most affected by censoring. The HEM leads to an elimination rate λ larger than HQL (difference approximatively two standard deviations). The standard errors of the estimates are artificially smaller for HQL. In order to highlight the influence of censoring on parameter estimates we conducted a second analysis of ACTG 315, using a censoring limit of 500 copies/mL (in HQL the censored data are replaced by 250 copies/mL). Overall, 9, 12 and 12 subjects had respectively one, two and more than two HIV RNA values censored at 500 copies/mL. The results of this analysis are included in Table 2, and the findings are similar to those in Table 1, with a more clear difference in the estimation of λ. Comparing λ between Table 1 and Table 2, the higher QL for censoring leads to lower values of λ. Heuristically, these values are biased downward, since more information for estimating λ is lost to censoring. The bias is larger for HQL. The higher standard deviation for λ for HEM compared to HQL reflects the adjustment for the uncertainty due to censoring.

Figure 2 displays fitted curves for four subjects, using HEM and HQL algorithms, based on QL of 500 copies/mL. The plots illustrate that the estimated decline in the second phase is steeper for HEM. As noted by Wu and Ding (1999), δ and λ represent the population-level elimination (turnover) rates for the two CD4 cell components (*e.g.*, productively infected cells and long-lived and/or latently infected cells respectively), and are important in understanding the pathogenesis of HIV-1 infection. They correspond to half-lives of $\ln(2)/\delta$ and $\ln(2)/\lambda$ respectively. For ACTG 315 data, the second CD4 cell component has an estimated half-life of 22.1 days (95% CI 17.3–30.3 days) using HEM, and 27.0 days (95% CI 22.1–34.7 days) using HQL. The program took 102 seconds on a Pentium M processor at 1.70 GHz. We used R version 1.9.0.

5 Hybrid EM versus Multiple Imputation: a simulation study

We compared the behavior and performance of the HEM with a Multiple Imputation method (Fitzgerald *et al.*, 2002) and the HQL algorithm via statistical simulation. Since the MLE for complete data NLME is only approximated, and potentially biased (Demidenko, 1997), we also include the complete data results (NLME), in order to separate the effect of censoring from the effect of the NLME approximation and finite sample bias. We used a simplified two-component exponential model (14). Noting that $P_{0i} = P_{1i} + P_{2i}$ is the expected baseline viral load for subject i (*i.e.* at $t = 0$), we write $P_{1i} = (1 + e^{\tau})^{-1}P_{0i}$, $P_{2i} = e^{\tau}(1 + e^{\tau})^{-1}P_{0i}$, and assume that the random effects are

$$(\ln P_{0i}, \lambda_i)^{\top} \overset{\text{iid}}{\sim} N((\ln P_0, \lambda)^{\top}, \sigma^2 D)$$

and the fixed effects are $\beta = (\ln P_0, \tau, \delta, \lambda)$. Note that the first component elimination rate δ is common to all subjects, $\delta_i = \delta$. The errors are independent $e_{ij} \sim N(0, \sigma^2)$. Fully observed data were obtained, which were subsequently censored at 100 copies/ml for the censored-data methods.

The simulation results are based on 500 simulated data sets. Parameter values β were chosen similar to the results of section 4 and of Wu and Ding (1999), but with a larger value of λ in order to avoid convergence problems: $\beta = (11.6, -3.6, 2.8 \text{ weeks}^{-1}, 0.35 \text{ weeks}^{-1})$, $\sigma^2 = 0.07$, The matrix D had elements $D_{11} = 2.25$, $D_{12} = D_{21} = 0.147$, $D_{22} = 0.02$. Each data set had 50 subjects, with 8 observations per subject and a follow-up of 12 weeks. Sixteen % of all observations and 38% of the observations after at least three weeks of follow-up were censored. At the estimation stage, the MI results are based on N = 5 imputations, with K = 50 iterations each. HQL used NLME with censored data replaced by 50 copies/mL. (An alternative approach to HQL would be to only replace the first censored observation by 50 copies/mL and discard the rest.) The simulation study was done in SAS.

The four methods were compared based on relative bias $E(\hat{\theta}-\theta)/|\theta|$, coverage probability for confidence intervals $Pr(\theta \in \hat{I})$, and relative root-mean-squared error $E^{1/2}[(\hat{\theta} - \theta)/\theta]^2$. (The parameter of interest θ is estimated by $\hat{\theta}$ and the confidence interval $\hat{I}$.) Table 3 presents the average fixed effects estimates and their simulation-based variance. Table 4 contains the relative bias and relative root mean squared error of the four methods, and Table 5 — the coverage probabilities. Finally, Table 6 contains the average estimates and relative bias for the variance components D and σ^2.

As expected, among the three methods dealing with censored data HEM performed best, followed by MI, while HQL gave generally poor results. The poorest performance in terms of both relative bias and coverage was obtained for λ, since the second turnover rate is most affected by censoring. There was a 37% bias in estimates using HQL; this bias fell to 4% for MI, and to less than 1% for HEM. The HEM estimate of λ had very good coverage, 93% for the 95% confidence interval, whereas MI had a coverage of 88%, and HQL had 0% coverage. Both MI and HEM algorithms account for censoring in the variance of the parameter estimates. HEM inflated estimates of the variance of $\ln P_0$, δ, λ and τ by 0%, 5%, 50% and 14% respectively. Alternatively, for each component of β we can compute the relative loss of information due to censoring, $1 - \text{Var}(\hat{\theta}_{\text{NLME}})/\text{Var}(\hat{\theta}_{\text{HEM}})$. For the four parameters, this is respectively 0%, 5%, 33%, and 12%. Note that one third of the information for estimating λ is lost due to censoring at QL = 100 copies/mL. Figure 3 presents the kernel density estimates for λ based on the simulations, and illustrates the bias and underestimation of variance for the

HQL estimates, the excellent performance of HEM and the very good results of MI. Similar results for the first turnover rate δ are presented in Figure 4.

Regarding the covariance parameters D_{ij} and σ^2 in Table 6, NLME had minimal bias; HQL performed poorly for all estimates. MI and HEM performed well for D_{11}, with a relative bias no greater than 1%. The three methods based on censored data underestimated D_{22}, with a bias of −6.5% for HEM, −21% for MI and −61% for HQL. The true correlation between random effects was 0.70. The average correlation for NLME, HQL, MI, and HEM was 0.70, −0.13, 0.73 and 0.66 respectively. The large underestimation of the variability in the λ_i random effect and of the correlation of the random effects when using HQL is expected, because this algorithm sets all censored values to a constant and so underestimates the heterogeneity in the decay rate. This underlines once again the need for appropriate adjustment for censoring.

6 Discussion

In this paper we have developed a hybrid EM algorithm for NLME models with censored response, and have compared its performance to multiple imputation and ad-hoc methods. The analyses have shown that even for relatively low levels of censoring the inference for certain model parameters may be severely biased if censoring is ignored. Multiple imputation may be implemented with existing NLME software and yields good working precision. However, it requires programming effort and statistical sophistication on the part of the user. The proposed HEM works in general situations with minimal user input and provides best precision and ease of use, at an affordable computation cost.

Demidenko (1997) found in a particular NLME case that the Lindstrom and Bates (1990) NLME approximation was asymptotically biased if the number of observations per subject n_{ij} is bounded. However, for our simulated NLME data examples, with moderate samples $n_{ij} = 8$, $m = 50$, the bias was negligible. This is in agreement with earlier results of Pinheiro and Bates (1995). Pinheiro and Bates (1995) also comment that the estimates of fixed effects and the estimates of covariance parameters appeared independent. Unlike the situation for linear mixed effects (Pinheiro, 1994), this independence has not been proven. Our simulation results suggest that this independence could still hold for NLME. However, for censored response NLME we found a significant correlation between estimates of fixed effects and covariance parameters. The correlation between the estimate of the fixed effect λ and the estimated variance of the random effect λ_i on the simulated data was −0.44 and −0.25 for HEM and MI algorithms respectively, versus −0.005 for NLME.

Currently our program computes the MLE for independent errors and for unstructured or diagonal random effects variance Ψ. While these are the most common situations, extensions are possible following the methodology presented here. These include, as in Pinheiro and Bates (2000), several levels of grouping; more complex structures for the diagonal matrix; or errors with serial or spatial correlation errors.

While we treated the censored observations as missing data, an alternative approach would be to maximize directly the censored data likelihood. Yet another approach is a Bayesian analysis using Markov chain Monte Carlo (Gilks *et al.*, 1996; Wakefield, 1996; Schafer, 1998), where the parameter estimates are based on samples from the posterior distribution.

The self-monitoring method for our algorithm offers a general stopping and monitoring rule which can be used in any MCEM setting. We note that the hybrid approach, using numerical integration at the E-step for clusters with one or two censored observations is speeding up the algorithm considerably. In our experience we found that Monte Carlo simulation is a necessary evil rather than an ideal solution to this problem. The good news is that through efficient

implementation of the M-step and sampling of the E-step our software provides solutions with good precision in real time.

Acknowledgments

The authors were partially supported for this work by NIH grants AI-28076-09, AI-38855, AI-51164, AI-51951, and AI-43638.

References

Bates, DM.; Pinheiro, JC. Tech rep. Bell Labs; New Jersey: 1997. Computational methods for multilevel modelling.

Bates, DM.; Watts, DG. Nonlinear Regression Analysis and Its Applications. Wiley; New York: 1988.

Beal SL. Ways to fit a PK model with some data below the quantification limit. Journal of Pharmacokinetics and Pharmacodynamics 2001;28:481–504. [PubMed: 11768292]

Booth JG, Hobert JP. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. Journal of the Royal Statistical Society, Series B 1999;61:265–285.

Chan KS, Ledolter J. Monte Carlo EM estimation of time series models involving counts. Journal of the American Statistical Society 1995;90(429):242–252.

Davidian, M.; Giltinan, DM. Nonlinear Models for Repeated Measurement Data. Chapman & Hall; 1995.

Demidenko, E. Asymptotic properties of nonlinear mixed-effects models. In: Gregoire, TG.; Brillinger, DR.; Diggle, PJ.; Russek-Cohen, E., editors. Modelling Longitudinal and Spatially Correlated Data: Methods, Applications, and Future Directions Lecture Notes in Statistics. Vol. 122. New York: Springer-Verlag; 1997. p. 49-62.

Diggle, PJ.; Heagerty, P.; Liang, KY.; Zeger, SL. Analysis of Longitudinal Data. Vol. 2nd. Oxford University Press; 2002.

Fitzgerald, A. Ph D thesis. Harvard University; Boston, Massachussets: 2000. Nonlinear mixed effects analysis for censored response.

Fitzgerald AP, DeGruttola VG, Vaida F. Modeling viral rebound using non-linear mixed effects models. Statistics in Medicine 2002;21:2093–2108. [PubMed: 12111889]

Genz A. Numerical computation of multivariate normal probabilities. Journal of Computational and Graphical Statistics 1992;1:141–150.

Gilks, WR.; Richardson, S.; Spiegelhalter, DJ., editors. Markov Chain Monte Carlo in Practice. Chapman and Hall; London: 1996.

Hughes JP. Mixed effects models with censored data with applications to HIV RNA levels. Biometrics 1999;55:625–629. [PubMed: 11318225]

Kiuchi AS, Hartigan JA, Holford TR, Rubinstein P, Stevens CE. Change points in the series of T4 counts prior to AIDS. Biometrics 1995;51:236–248. [PubMed: 7766779]

Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics 1982;38:963–974. [PubMed: 7168798]

Lederman MM, Connick E, Landay A, Kuritzkes DR, Spritzler J, St Clair M, Kotzin BL, Fox L, Chiozzi MH, Leonard JM, Rousseau F, Wade M, D'Arc Roe J, Martinez A, Kessier H. Immunological responses associated with 12 weeks of combination antiretroviral therapy consisting of Zidovudine, Lamivudine, and Ritonavir: Results of AIDS clinical trial group protocol 315. Journal of Infectious Diseases 1998;178:70–79. [PubMed: 9652425]

Lindstrom MJ, Bates DM. Nonlinear mixed effects models for repeated measures data. Biometrics 1990;46:673–687. [PubMed: 2242409]

Louis TA. Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society, Series B 1982;44:226–233.

Maddala, GS. Limited-dependent and Qualitive Variables in Econometrics. Cambridge University Press; 1996.

Meng XL, van Dyk D. The EM algorithm – an old folk-song sung to a fast new tune. Journal of the Royal Statistical Society, Series B 1997;59(3):511–567.

Orchard T, Woodbury MA. A missing information principle, theory and application. Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability 1972;1:679–715.

Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD. HIV-1 dynamics in vivo: Virion clearance rate, infected cell life-span, and viral generation time. Science 1996;271:1582–1586. [PubMed: 8599114]

Pinheiro, JC. Ph D thesis. University of Wisconsin; Madison, WI: 1994. Topics in Mixed-Effects Models.

Pinheiro JC, Bates DM. Approximations to the log-likelihood function in the nonlinear mixed-effects model. Journal of Computational and Graphical Statistics 1995;4:12–35.

Pinheiro, JC.; Bates, DM. Mixed-Effects Models in S and S-PLUS. Springer; New-York: 2000.

Rubin DB. Multiple imputation after 18+ years (with discussion). Journal of the American Statistical Association 1996;91:473–520.

Schafer, JL. Tech rep. Pennsylvania State University, Department of Statistics; 1998. Some improved procedures for linear mixed model.

Sheiner LB, Beal SL. Evaluation of methods for estimating population pharmokinetics. I. Michaelis-Menton Model: Routine clinical pharacokinetics data. Journal of Pharmacokinetics and Biopharmaceutics 1980;8:673–687.

Vaida F, Meng XL. Two slice-EM algorithms for fitting generalized linear mixed models with binary response. Statistical Modelling 2005;5:229–242.

Vonesh, EF.; Chinchilli, VM. Linear and Nonlinear Models for the Analysis of Repeated Measurements. Marcel Dekker; 1997.

Wakefield J. The Bayesian analysis of population pharmacokinetic models. Journal of the American Statistical Association 1996;91:62–75.

Wolfinger R. Laplace's approximation for nonlinear mixed models. Biometrika 1993;80:791–795.

Wolfinger, R. Tech Rep. SAS Institute Inc; Cary, NC: 1999. Fitting nonlinear mixed models with the new NLMIXED procedure; p. 287

Wu H, Ding A. Population HIV-1 dynamics in vivo: Applicable models and inferential tools for virological data from AIDS clinical trials. Biometrics 1999;55:410–418. [PubMed: 11318194]

Wu L. A joint model for nonlinear mixed-effects models with censoring and covariates measured with error, with application to AIDS studies. Journal of the American Statistical Society 2002;97:955–964.

Wu L. Simultaneous inference for longitudinal data with detection limits and covariates measured with errors, with application to AIDS studies. Statistics in Medicine 2004;23:1715–1731. [PubMed: 15160404]

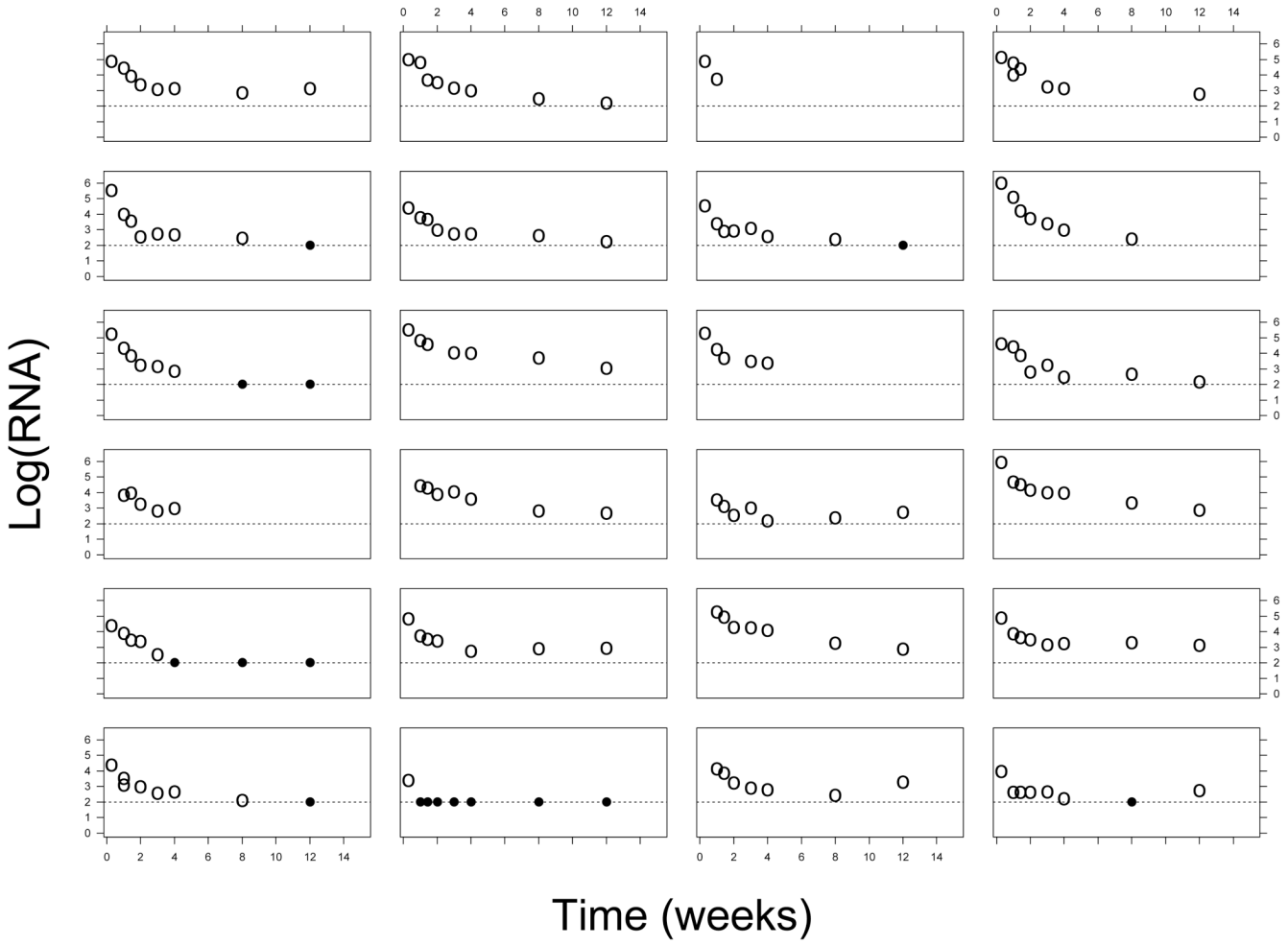

Figure 1.
ACTG 315 data: $\log_{10}$ HIV-1 RNA for the first 12 weeks for 24 subjects. Observed responses (◦) are shown along with censored readings (•); Lower QL = 100 copies/mL.

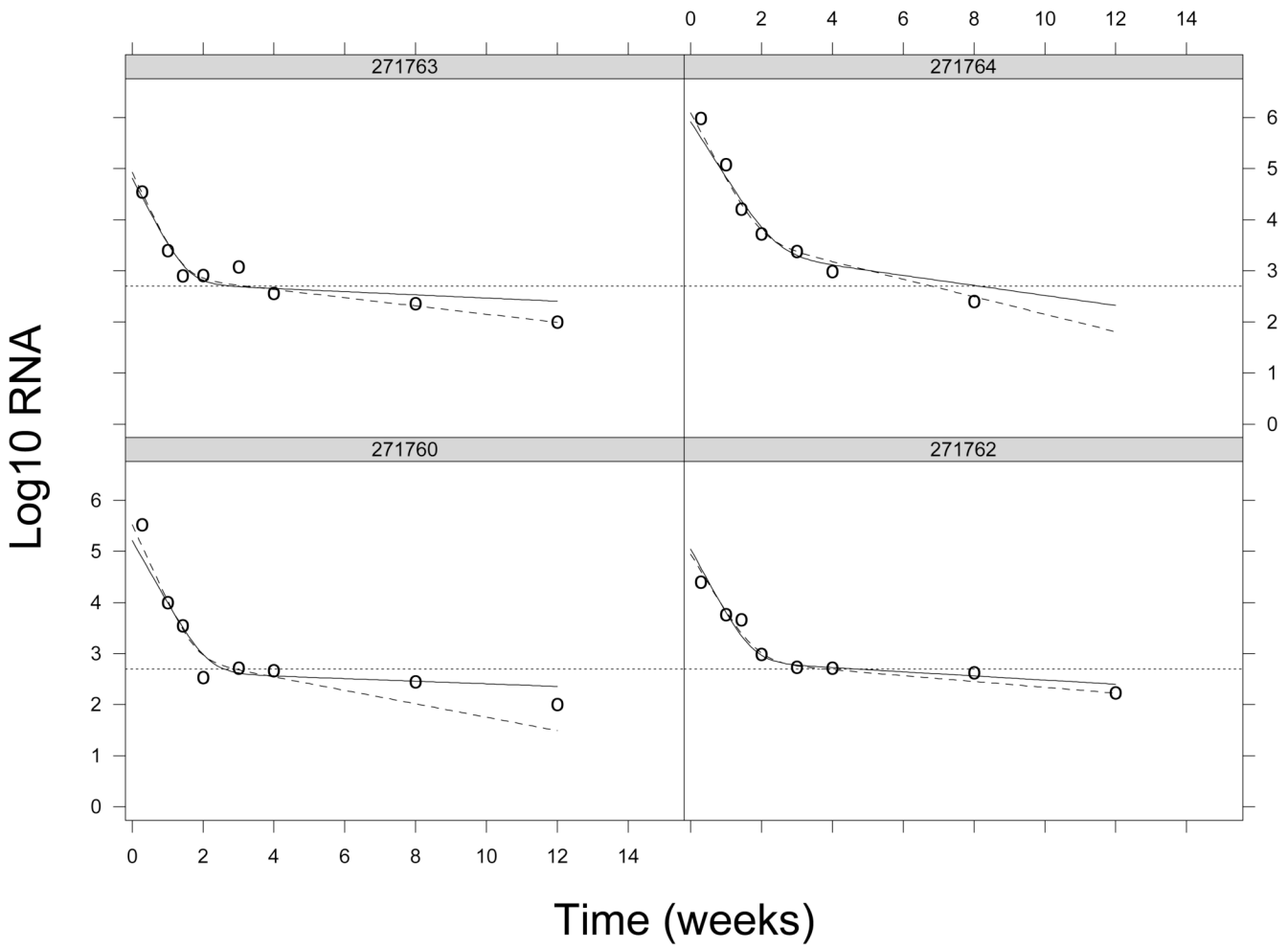

Figure 2.
ACTG 315 data: HIV-1 RNA response and estimated subject-specific response for four subjects, based on QL = 500 copies/mL (dotted line). The two fitting methods are HEM (dashed line), and HQL (solid line).

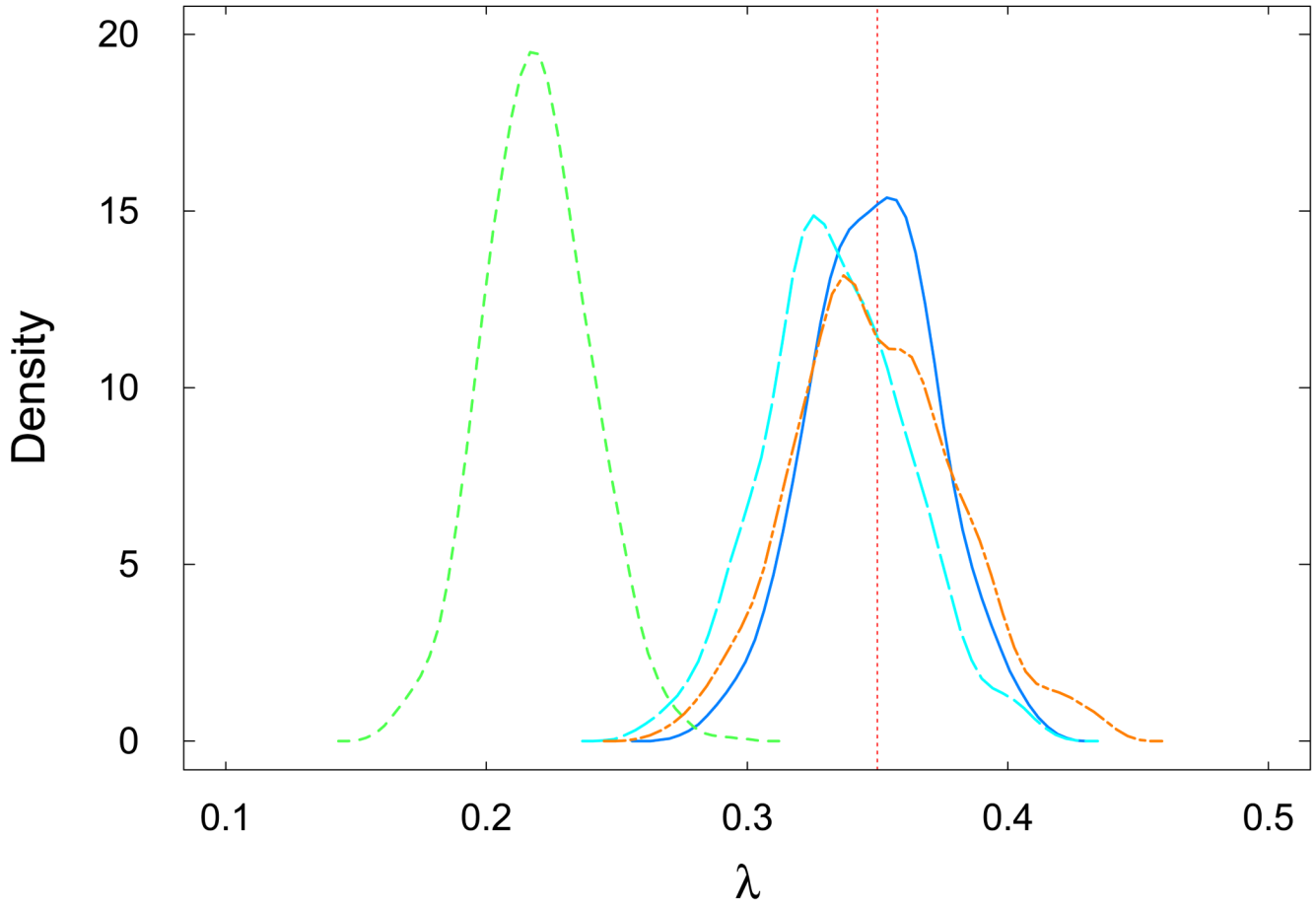

Figure 3.
Simulation study: estimates of the fixed effect λ, using NLME (—), HEM (– · –), MI (– –), and HQL (- -). Vertical line represents the true value.

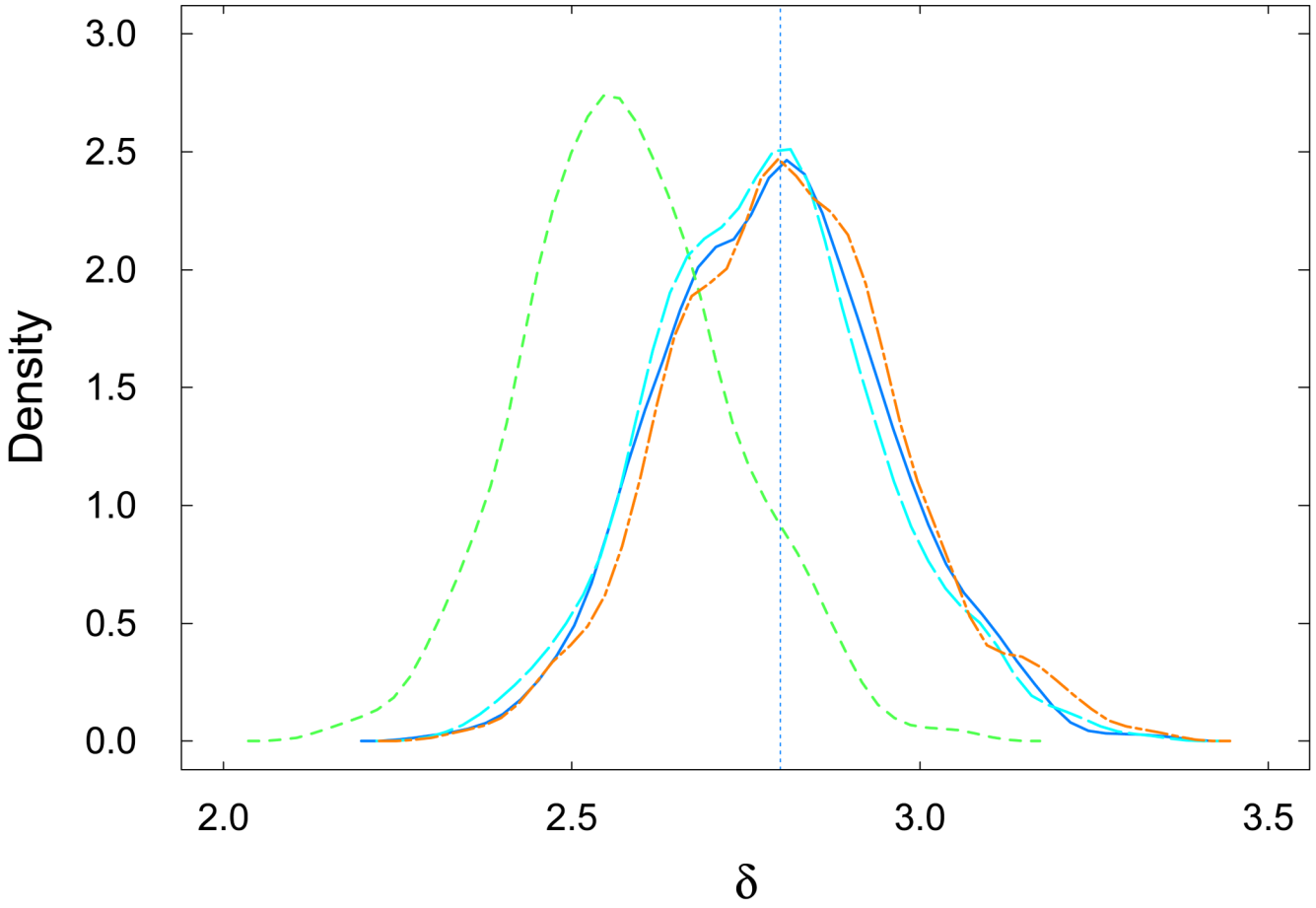

Figure 4.
Simulation study: estimates of the fixed effect β, using NLME (—), HEM (– · –), MI (– –), and HQL (- -). Vertical line represents the true value.

Table 1

ACTG 315 results, QL = 100 copies/mL. HEM: NLME for censored data using HEM. HQL: NLME, censored data replaced by 50 copies/mL.

	HEM		HQL	
	Estimate	**SE**	**Estimate**	**SE**
δ (weeks^{-1})	2.392	0.108	2.375	0.104
λ (weeks^{-1})	0.2444	0.0306	0.2148	0.0253
$\ln P_1$	11.55	0.24	11.55	0.23
$\ln P_2$	7.783	0.265	7.685	0.259
σ	0.2570		0.2604	

Table 2

ACTG 315 results, QL = 500 copies/mL. *HEM*: NLME for censored data using HEM. *HQL*: NLME, censored data replaced by 250 copies/mL.

	HEM		HQL	
	Estimate	**SE**	**Estimate**	**SE**
δ (weeks^{-1})	2.344	0.103	2.666	0.099
λ (weeks^{-1})	0.1990	0.0321	0.1327	0.0202
$\ln P_1$	11.55	0.23	11.52	0.23
$\ln P_2$	7.627	0.281	7.434	0.236
σ	0.2505		0.2494	

Table 3

Estimated fixed effects from the simulation study (mean, standard deviation).

	$\ln P_0$	δ	λ	τ
True value	11.60	2.80	0.35	-3.50
HQL	11.52 (0.23)	2.58 (0.15)	0.22 (0.02)	-3.89 (0.14)
MI	11.59 (0.23)	2.78 (0.16)	0.33 (0.03)	-3.55 (0.14)
HEM	11.60 (0.23)	2.81 (0.16)	0.35 (0.03)	-3.51 (0.14)
NLME	11.59 (0.23)	2.80 (0.16)	0.35 (0.02)	-3.51 (0.13)

Table 4

Relative bias (%) and relative root mean square error (in parentheses, %), for fixed effects estimates from the simulation study.

	$\ln P_0$	δ	λ	τ
HQL	-0.73 (2.2)	-7.87 (9.5)	-37.3 (37.7)	-11.1 (11.8)
MI	-0.08 (2.0)	-0.75 (5.9)	-4.51 (9.0)	-1.41 (4.2)
HEM	0.01 (2.0)	0.32 (5.9)	-0.28 (8.8)	-0.01 (4.0)
NLME	-0.05 (2.0)	-0.25 (5.8)	-0.40 (6.8)	0.24 (3.7)

Table 5

Coverage probabilities for 95% confidence intervals on the fixed effects from the simulation study.

	$\ln P_0$	δ	λ	τ
HQL	0.95	0.62	0.00	0.21
MI	0.96	0.93	0.88	0.95
HEM	0.96	0.94	0.93	0.94
NLME	0.96	0.94	0.94	0.95

Table 6

Estimated Covariance parameters and relative bias (%, in parentheses) from the simulation study.

	D_{11}	D_{22}	D_{12}	σ^2
True values	2.25	0.020	0.147	0.070
HQL	2.65 (18)	0.001 (-61)	-0.019 (-113)	0.075 (7.5)
MI	2.26 (0.6)	0.015 (-22)	0.135 (-8.5)	0.068 (-2.4)
HEM	2.23 (-1.0)	0.021 (-6.5)	0.141 (-4.2)	0.068 (-3.3)
NLME	2.24 (-0.4)	0.020 (-0.3)	0.147 (0.3)	0.070 (-0.6)