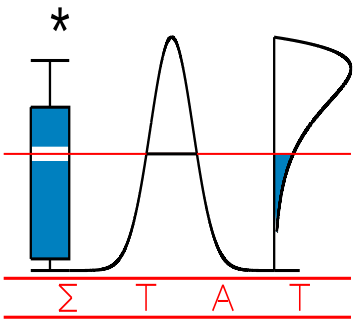
T E C H N I C A L

R E P O R T

0650

A LATENT-CLASS MIXTURE MODEL FOR INCOMPLETE LONGITUDINAL GAUSSIAN DATA

BEUNCKENS C., MOLENBERGHS G., VERBEKE G., and C. MALLINCKRODT



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

A Latent-Class Mixture Model For Incomplete Longitudinal Gaussian Data

Caroline Beunckens,[1,*] **Geert Molenberghs,**[1]
Geert Verbeke,[2] **and Craig Mallinckrodt**[3]

[1] Center for Statistics, Hasselt University, Agoralaan 1, 3590 Diepenbeek, Belgium.
[2] Biostatistical Centre, Catholic University of Leuven,
Kapucijnenvoer 35, 3000 Leuven, Belgium.
[3] Eli Lilly & Company, Lilly Corporate Center, Indianapolis, IN 46285, U.S.A.
[*] *email:* caroline.beunckens@uhasselt.be

Summary. In the analyses of incomplete longitudinal clinical trial data, there has been a shift, away from simple methods that are valid only if the data are missing completely at random (MCAR), to more principled ignorable analyses, which are valid under the less restrictive missing at random (MAR) assumption. The availability of the necessary standard statistical software nowadays allows for such analyses in practice. While the possibility of data missing not at random (MNAR) cannot be ruled out, it is argued that analyses valid under MNAR are not well suited for the primary analysis in clinical trials. Rather than either forgetting about or blindly shifting to an MNAR framework, the optimal place for MNAR analyses is within a sensitivity analysis context. One such route for sensitivity analysis is to consider, next to selection models, pattern-mixture models or shared-parameter models. The latter can also be extended to a latent-class mixture model, the route taken in this paper. The so-obtained flexible model is submitted to the test in simulations and applied to data from a depression trial.

Key Words: Latent class, Nonrandom missingness, Random effect, Shared parameter

1 Introduction

Data arising from studies with observations made repeatedly over time are often prone to incompleteness. In the context of such longitudinal studies, missingness predominantly occurs in the form of dropouts, in which subjects fail to complete the study for one reason or another.

The nature of the dropout mechanism can affect the results from analyses of incomplete data. Since one can never be certain about the dropout mechanism, certain assumptions have to be made. When referring to the missingness process, we will use the terminology introduced by Rubin (1976). A non-response process is said to be *missing completely at random* (MCAR) if the missingness is independent of both unobserved and observed data, and *missing at random* (MAR) if, conditional on the observed data, the missingness is independent of the unobserved measurements. A process that is neither MCAR nor MAR is termed *non-random* (MNAR). In the context of likelihood inference, and when the parameters describing the measurement process are functionally independent of the parameters describing the missingness process, MCAR and MAR are *ignorable*, in which case the missingness process can be ignored when interest is in inference for the longitudinal process only, whereas an MNAR process is non-ignorable. In contrast, for frequentist inference, the stronger condition of MCAR is required to ensure ignorability.

Simple methods to tackle incomplete longitudinal data, such as complete case analysis, last observation carried forward analysis, and other forms of single imputation, have been popular. However, a major drawback is that such methods are valid only under the restrictive MCAR assumption or even more stringent conditions. Hence, the primary analyses should move from such simple methods to an analysis valid under the MAR assumption, which is not more difficult to conduct (Molenberghs et al., 2004; Jansen et al., 2006). For example, likelihood-based analyses using linear mixed or generalized linear mixed models are valid under MAR, and can be conducted using standard statistical software. Further, since non-random methods allow the missingness to depend on the unobserved or missing values, it is clear that the MNAR assumption is not verifiable (Laird, 1994; Molenberghs, Kenward and

Lesaffre, 1997). Therefore, fitting a single MNAR model will not be trustworthy. However, this does not mean we should ignore such models, rather we should use them in a sensitivity analysis framework.

The distinction between the three different missingness mechanisms, as described above, is made within the framework of *selection models*. Another form of sensitivity analysis, besides fitting several plausible M(N)AR models, consists of using pattern-mixture or shared-parameter models, which are introduced in the next section.

In this paper, we propose a so-called latent-class mixture model, bringing together features of the selection, pattern-mixture, and shared-parameter model frameworks. Precisely, information from the location and evolution of the response profiles, a selection model concept, and from the dropout patterns, a pattern-mixture idea, is used simultaneously to define latent groups and variables, a shared-parameter feature. This approach has a number of appealing features. First, it allows for using the information in a more symmetric and therefore more elegant way. Second, apart from providing a more flexible modeling tool, the new framework is ideally suited to be used as a sensitivity analysis instrument. Third, a strong added advantage over existing methods is that we now will be able to classify subjects into latent groups. While this has to be done with due caution, it can enhance substantive knowledge and generate hypotheses for further research. Fourth, while computational burden evidently increases, fitting the proposed method is remarkably stable and falls within acceptable time limits for applications of the type considered here and for simulations reported in this paper.

The reader should be aware that neither the proposed model nor any other alternative can be seen as a tool to, for example, definitively test the null hypothesis of MAR *versus* the MNAR alternative, a fact amply documented in the sensitivity analysis literature. This is why the method's use lies predominantly within the sensitivity analysis context. Such a sensitivity analysis is clearly useful when the more elaborate model modifies the results from the simpler alternative. However, even when it confirms earlier results, as will be the case in our data analysis, it will typically increase confidence in the conclusions reached.

The general latent-class mixture model is presented in Section 2. To show the performance of this model, a simulations study is considered in Section 5. Finally, in Section 6, data from a depression clinical trial are analyzed using a latent-class mixture model within a sensitivity analysis.

2 Latent-Class Mixture Models

Let the random variable Y_{ij} denote the response of interest, for the ith individual, designed to be measured at time t_{ij}, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$. The outcomes can be grouped into a vector $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$. In addition, define a dropout indicator D_i for the occasion at which dropout occurs, i.e., Y_{d_i} is the first missing value, and apply the convention that $D_i = n_i + 1$ for a complete sequence. It is often convenient to split the vector $\boldsymbol{Y}_i$ into observed ($\boldsymbol{Y}_i^o$) and missing ($\boldsymbol{Y}_i^m$) components, respectively.

In principle, one would like to consider the density of the full data $f(\boldsymbol{y}_i, d_i|\boldsymbol{\theta}, \boldsymbol{\psi})$, where the parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ describe the measurement and missingness processes, respectively. Covariates are assumed to be measured, but have been suppressed from notation for simplicity.

This full density function can be factorized in different ways, each leading to a different framework. The *selection model* framework is based on the following factorization (Rubin, 1976; Little and Rubin, 1987):

$$f(\boldsymbol{y}_i, d_i|\boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i|\boldsymbol{\theta}) f(d_i|\boldsymbol{y}_i, \boldsymbol{\psi}).$$

The first factor is the marginal density of the measurement process and the second one is the density of the missingness process, conditional on the outcomes. As an alternative, one can consider so-called *pattern-mixture models* (Little, 1993, 1994) using the reversed factorization

$$f(\boldsymbol{y}_i, d_i|\boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i|d_i, \boldsymbol{\theta}) f(d_i|\boldsymbol{\psi}).$$

This can be seen as a mixture of different populations, characterized by the observed pattern of missingness.

Instead of using the selection modelling or pattern-mixture modelling framework, the measurement and the dropout process can be jointly modelled by using a *shared-parameter model* as it is introduced in Wu and Carrol (1988), Ten Have et al. (1998), Wu and Bailey (1989), Mori, Woodworth and Woolson (1992), Follmann and Wu (1995), and Little (1995). These methods assume there exists a vector of random effects $\boldsymbol{b}_i$, conditional upon which the measurement and dropout processes are independent. This shared-parameter model can be formulated by the following factorization

$$f(\boldsymbol{y}_i, d_i|\boldsymbol{b}_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\theta}) f(d_i|\boldsymbol{b}_i, \boldsymbol{\psi}).$$

We propose an extension of this shared parameter model capturing possible heterogeneity between the subjects, which is not measured through a covariate, but rather through a latent variable. We call this model a *latent-class mixture model*. Next to one or more random effects, or so-called shared parameters, $\boldsymbol{b}_i$, such a model contains a latent variable, $\boldsymbol{Q}_i$, dividing the population in g subgroups. This latent variable is a vector of group indicators $\boldsymbol{Q}_i = (Q_{i1}, \ldots, Q_{ig})$, defined as $Q_{ik} = 1$, if subject i belongs to group k, and 0 otherwise. The measurement process as well as the dropout process depend on this latent variable, not only directly, but also through the subject-specific effects $\boldsymbol{b}_i$. The distribution of $\boldsymbol{Q}_i$ is multinomial and defined by $P(Q_{ik} = 1) = \pi_k$, where k ranges from 1 to g and π_k denotes the group or component probability. Note that the component probabilities are restricted through $\sum_{k=1}^{g} \pi_k = 1$. In what follows, π_k will also be called the prior probability for any observation to belong to the kth component of the mixture.

The measurement process will be modelled by a heterogeneity linear mixed model proposed by Verbeke and Lesaffre (1996) and also described by Verbeke and Molenberghs (2000, Chapter 12), i.e.,

$$\boldsymbol{Y}_i|q_{ik} = 1, \boldsymbol{b}_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta}_k + \boldsymbol{Z}_i\boldsymbol{b}_i, \boldsymbol{\Sigma}_i^{(k)}),$$

where $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are design matrices, $\boldsymbol{\beta}_k$ are fixed effects, possibly depending on the group components, $\boldsymbol{b}_i$ denote the shared parameters, following a mixture of g normal distributions with mean vectors $\boldsymbol{\mu}_k$ and covariance matrices $\boldsymbol{D}_k$, i.e.,

$$\boldsymbol{b}_i|q_{ik} = 1 \sim N(\boldsymbol{\mu}_k, \boldsymbol{D}_k),$$

and thus

$$\boldsymbol{b}_i \sim \sum_{k=1}^{g} \pi_k N(\boldsymbol{\mu}_k, \boldsymbol{D}_k).$$

The measurement error terms $\boldsymbol{\varepsilon}_i$ follow a normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}_i^{(k)}$ and are independent of the shared parameters. The mean and the variance of $\boldsymbol{Y}_i$ can be derived as:

$$E(\boldsymbol{Y}_i) = \boldsymbol{X}_i \sum_{k=1}^{g} \pi_k \boldsymbol{\beta}_k + \boldsymbol{Z}_i \sum_{k=1}^{g} \pi_k \boldsymbol{\mu}_k, \tag{1}$$

$$\mathrm{Var}(\boldsymbol{Y}_i) = \boldsymbol{Z}_i' \left[\sum_{k=1}^{g} \pi_k \boldsymbol{\mu}_k^2 - \left(\sum_{k=1}^{g} \pi_k \boldsymbol{\mu}_k \right)^2 + \sum_{k=1}^{g} \pi_k \boldsymbol{D}_k \right] \boldsymbol{Z}_i + \sum_{k=1}^{g} \pi_k \boldsymbol{\Sigma}_i^{(k)}. \tag{2}$$

Further, we have to assume that the shared effects are 'calibrated', i.e., $\sum_{k=1}^{g} \pi_k \boldsymbol{\mu}_k = \boldsymbol{0}$, then (1) and (2) simplify to:

$$E(\boldsymbol{Y}_i) = \boldsymbol{X}_i \sum_{k=1}^{g} \pi_k \boldsymbol{\beta}_k,$$

$$\mathrm{Var}(\boldsymbol{Y}_i) = \boldsymbol{Z}_i' \left[\sum_{k=1}^{g} \pi_k \boldsymbol{\mu}_k^2 + \sum_{k=1}^{g} \pi_k \boldsymbol{D}_k \right] \boldsymbol{Z}_i + \sum_{k=1}^{g} \pi_k \boldsymbol{\Sigma}_i^{(k)}.$$

Assuming that the first measurement Y_{i1} is obtained for every subject in the study, the model for the dropout process is based on a logistic regression for the probability of dropout at occasion j, given the subject was still in the study up to occasion j, given the random effects $\boldsymbol{b}_i$, and given that the subject belonged to the kth component of the mixture. We denote this probability by $g_{ij}(\boldsymbol{w}_{ij}, \boldsymbol{b}_i, q_{ik})$, in which $\boldsymbol{w}_{ij}$ is a vector containing all relevant covariates: $g_{ij}(\boldsymbol{w}_{ij}, \boldsymbol{b}_i, q_{ik}) = P(D_i = j | D_i \geq j, \boldsymbol{w}_{ij}, \boldsymbol{b}_i, q_{ik} = 1)$. We then assume that $g_{ij}(\boldsymbol{w}_{ij}, \boldsymbol{b}_i, q_{ik})$ satisfies $\mathrm{logit}[g_{ij}(\boldsymbol{w}_{ij}, \boldsymbol{b}_i, q_{ik})] = \boldsymbol{w}_{ij}\boldsymbol{\gamma}_k + \boldsymbol{\lambda}\boldsymbol{b}_i$. Now, the joint likelihood of the measurement and dropout processes will take the form:

$$\begin{aligned} f(\boldsymbol{y}_i, d_i) &= \sum_{k=1}^{g} P(q_{ik} = 1) f(\boldsymbol{y}_i, d_i | q_{ik} = 1) \\ &= \sum_{k=1}^{g} \pi_k \int f(\boldsymbol{y}_i, d_i | q_{ik} = 1, \boldsymbol{b}_i) f_k(\boldsymbol{b}_i) d\boldsymbol{b}_i \\ &= \sum_{k=1}^{g} \pi_k \int f(\boldsymbol{y}_i | q_{ik} = 1, \boldsymbol{b}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i) f(d_i | q_{ik} = 1, \boldsymbol{b}_i, \boldsymbol{w}_i) f_k(\boldsymbol{b}_i) d\boldsymbol{b}_i, \end{aligned} \tag{3}$$

with $f(\boldsymbol{y}_i|q_{ik}=1,\boldsymbol{b}_i,\boldsymbol{X}_i,\boldsymbol{Z}_i)$ the density function of the normal distribution $N(\boldsymbol{X}_i\boldsymbol{\beta}_k + \boldsymbol{Z}_i\boldsymbol{b}_i, \boldsymbol{\Sigma}_i^{(k)})$, $f_k(\boldsymbol{b}_i)$ is the density function of $N(\boldsymbol{\mu}_k, \boldsymbol{D}_k)$, and

$$f(d_i|q_{ik}=1,\boldsymbol{b}_i,\boldsymbol{w}_i) = \begin{cases} g_{id_i}(\boldsymbol{w}_{id_i},\boldsymbol{b}_i,q_{ik}) \times \prod_{j=2}^{d_i-1}[1-g_{ij}(\boldsymbol{w}_{ij},\boldsymbol{b}_i,q_{ik})] & \text{if incomplete,} \\ \prod_{j=2}^{n_i}[1-g_{ij}(\boldsymbol{w}_{ij},\boldsymbol{b}_i,q_{ik})] & \text{if complete.} \end{cases}$$

Whereas selection models and pattern-mixture models derive from two different factorizations of the joint density of the measurement and dropout processes, the latent-class mixture model is based on assuming an additional latent structure. The selection model lends itself naturally to formulate such concepts as MAR and ignorability, even though they can be considered in the pattern-mixture framework as well (Molenberghs et al., 1998; Kenward, Molenberghs, and Thijs, 2003). In the pattern-mixture model, the observed dropout patterns are taken into account when modeling the measurement process. The latent-class mixture models modify this idea by grouping the subjects by means of a latent variable, thereby accounting for inter-group differences both in terms of their dropout pattern as well as their measurement profiles.

3 Likelihood Function and Estimation

Estimation of the unknown parameters in the latent-class mixture model described in the previous section will be based on the maximum likelihood principle. To this end, the likelihood function of the latent-class mixture model is formulated in Section 3.1. Since it would be very cumbersome to maximize this likelihood function analytically, the EM algorithm (Dempster, Laird and Rubin, 1977) is proposed as it is a practical tool for maximum likelihood estimation in the case of finite mixtures (Redner and Walker, 1984).

3.1 The Likelihood Function

Let $\boldsymbol{\pi}$ be the vector of component probabilities $\boldsymbol{\pi}' = (\pi_1, \ldots, \pi_g)$ and group all other unknown parameters of the measurement process in the vector $\boldsymbol{\theta}$, of the dropout process in $\boldsymbol{\psi}$, and

of the mixture distribution in $\boldsymbol{\alpha}$. If $\boldsymbol{\sigma}$ denotes the vector of covariance parameters of all $\boldsymbol{\Sigma}_i^{(k)}$, $\boldsymbol{\delta}$ the covariance parameters of all $\boldsymbol{D}_k$, $\boldsymbol{\mu}' = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g)$, and $\boldsymbol{\gamma}' = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_g)$, then $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\sigma})$, $\boldsymbol{\psi} = (\boldsymbol{\gamma}, \boldsymbol{\lambda})$ and $\boldsymbol{\alpha} = (\boldsymbol{\mu}, \boldsymbol{\delta})$. Now, the vector $\boldsymbol{\Omega}$ will be the vector containing all unknown parameters in the model, i.e., $\boldsymbol{\Omega}' = (\boldsymbol{\pi}', \boldsymbol{\theta}', \boldsymbol{\psi}', \boldsymbol{\alpha}')$.

Estimation and inference for the $\boldsymbol{\Omega}$ will now be based on the observed data likelihood, $L(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d})$, obtained by integrating the unobserved data out of the joint distribution of measurement and dropout process and expressed by:

$$\begin{aligned}
L(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d}) &= \prod_{i=1}^{N} f(\boldsymbol{y}_i^o, d_i|\boldsymbol{\Omega}) = \prod_{i=1}^{N} \int f(\boldsymbol{y}_i, d_i|\boldsymbol{\Omega})\, d\boldsymbol{y}_i^m \\
&= \prod_{i=1}^{N} \int \left\{ \sum_{k=1}^{g} \pi_k \int f(\boldsymbol{y}_i|\boldsymbol{\theta}, \boldsymbol{b}_i, q_{ik}=1) f(d_i|\boldsymbol{\psi}, \boldsymbol{b}_i, q_{ik}=1) f_k(\boldsymbol{b}_i|\boldsymbol{\alpha}) d\boldsymbol{b}_i \right\} d\boldsymbol{y}_i^m \\
&= \prod_{i=1}^{N} \sum_{k=1}^{g} \pi_k \int \left\{ \int f(\boldsymbol{y}_i|\boldsymbol{\theta}, \boldsymbol{b}_i, q_{ik}=1)\, d\boldsymbol{y}_i^m \right\} f(d_i|\boldsymbol{\psi}, \boldsymbol{b}_i, q_{ik}=1) f_k(\boldsymbol{b}_i|\boldsymbol{\alpha}) d\boldsymbol{b}_i \\
&= \prod_{i=1}^{N} \sum_{k=1}^{g} \pi_k \int f(\boldsymbol{y}_i^o|\boldsymbol{\theta}, \boldsymbol{b}_i, q_{ik}=1) f(d_i|\boldsymbol{\psi}, \boldsymbol{b}_i, q_{ik}=1) f_k(\boldsymbol{b}_i|\boldsymbol{\alpha}) d\boldsymbol{b}_i, \qquad (4)
\end{aligned}$$

where $\boldsymbol{y}^{o'} = (\boldsymbol{y}_1^o, \ldots, \boldsymbol{y}_N^o)$ is the vector containing all observed response values and $\boldsymbol{d} = (d_1, \ldots, d_N)$ is the vector of all values of the dropout indicator.

Note that this likelihood function is invariant under the $g!$ possible permutations of the parameters corresponding to each of the g mixture components. However, we can put some constraints on the parameters, such that this problem of lack of identifiability disappears. We will use the constraint suggested by Aitkin and Rubin (1985), $\pi_1 \geq \pi_2 \geq \ldots \geq \pi_g$.

More generally, identifiability is an important but tricky issue. While it has been settled in a number of relatively simple settings, such as a two-component mixture of normals, arguably such a treatment is next to impossible in such complicated settings as ours, where apart from latent classes, also latent variables (random effects), and latency due to missingness arises. The best, admittedly pragmatic, piece of advice is to consider a variety of slight variations to the target model. The likelihood values, parameter estimates, and the information matrices can then be studied in view of identifiability.

The log-likelihood function corresponding to likelihood function (4) is

$$\ell(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d}) = \sum_{i=1}^{N} \ln \left\{ \sum_{k=1}^{g} \pi_k \int f(\boldsymbol{y}_i^o|\boldsymbol{\theta}, \boldsymbol{b}_i, q_{ik}=1) f(d_i|\boldsymbol{\psi}, \boldsymbol{b}_i, q_{ik}=1) f_k(\boldsymbol{b}_i|\boldsymbol{\alpha}) d\boldsymbol{b}_i \right\}. \quad (5)$$

To maximize (5) with respect to $\boldsymbol{\Omega}$, we will need a numerical iterative procedure. The EM algorithm is designed for maximum likelihood estimation in situations with missing data (Dempster et al., 1977). Here, the underlying latent variable $\boldsymbol{Q}_i$, representing component membership, will be considered missing. Thus, the response vector $\boldsymbol{Y}_i^o$ and the dropout indicator D_i, together with the (unobserved) population indicators $\boldsymbol{Q}_i$ can be seen as the augmented data, whereas vectors $\boldsymbol{Y}_i^o$ and D_i alone are the observed data.

The likelihood function $L(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d})$ still corresponds to the incomplete data. Since the joint density of $\boldsymbol{Y}_i^o$, D_i and $\boldsymbol{Q}_i$ equals

$$\begin{aligned}
& f_i(\boldsymbol{y}_i^o, d_i, Q_{i1}=q_{i1}, \ldots, Q_{ig}=q_{ig}) \\
&= f_i(\boldsymbol{y}_i^o, d_i|Q_{i1}=q_{i1}, \ldots, Q_{ig}=q_{ig}) \times P(Q_{i1}=q_{i1}, \ldots, Q_{ig}=q_{ig}) \\
&= \left\{ \prod_{k=1}^{g} \left[f_{ik}(\boldsymbol{y}_i^o, d_i|\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\alpha}) \right]^{q_{ik}} \right\} \cdot \left\{ \prod_{k=1}^{g} \pi_k^{q_{ik}} \right\} \\
&= \prod_{k=1}^{g} \left[\pi_k f_{ik}(\boldsymbol{y}_i^o, d_i|\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\alpha}) \right]^{q_{ik}},
\end{aligned}$$

the joint likelihood $L(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d}, \boldsymbol{q})$ of the augmented data, i.e., the likelihood function that would have been obtained if the values $\boldsymbol{q}_i = (q_{i1}, \ldots, q_{ig})'$ of the population indicators $\boldsymbol{Q}_i$ had been observed, will be

$$L(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d}, \boldsymbol{q}) = \prod_{i=1}^{N} \prod_{k=1}^{g} \left[\pi_k f_{ik}(\boldsymbol{y}_i^o, d_i|\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\alpha}) \right]^{q_{ik}}, \quad (6)$$

with $\boldsymbol{q} = (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n)'$ the vector of all hypothetically observed population indicators. The log-likelihood function corresponding to likelihood function (6) will be of the form

$$\ell(\boldsymbol{\Omega}|\boldsymbol{y}, \boldsymbol{d}, \boldsymbol{q}) = \sum_{i=1}^{N} \sum_{k=1}^{g} q_{ik} \left\{ \ln \pi_k + \ln f_{ik}(\boldsymbol{y}_i^o, d_i|\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\alpha}) \right\}. \quad (7)$$

3.2 Estimation Using The EM Algorithm

Maximizing $\ell(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d}, \boldsymbol{q})$ will be analytically and computationally easier than maximizing the log-likelihood $\ell(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d})$. However, the estimates obtained from maximizing $\ell(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d}, \boldsymbol{q})$ with respect to $\boldsymbol{\Omega}$, will depend on the unobserved indicators $\boldsymbol{q}$. Therefore, the EM algorithm will be used, since this will maximize the expected value of $\ell(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d}, \boldsymbol{q})$ with respect to $\boldsymbol{\Omega}$, where the expectation is taken over all unobserved $\boldsymbol{q}$, i.e., $E[\ell(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d}, \boldsymbol{Q})|\boldsymbol{y}, \boldsymbol{d}]$. This conditional expectation of $\ell(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d}, \boldsymbol{q})$ given $\boldsymbol{y}^o$ and $\boldsymbol{d}$, is calculated within the expectation (E) step of each iteration of the EM algorithm. In the maximization (M) step of the EM algorithm the expected log-likelihood function obtained from the E step is then maximized. We will denote the expected log-likelihood function by $\mathcal{O}$ and call it the objective function. The EM algorithm is an iterative procedure, i.e., it starts from an initial value $\boldsymbol{\Omega}^{(0)}$ for $\boldsymbol{\Omega}$, and then constructs a series of estimates $\boldsymbol{\Omega}^{(t)}$, which converges to the maximum likelihood estimator $\widehat{\boldsymbol{\Omega}}$ of $\boldsymbol{\Omega}$. Initial values can be obtained from considering separate models for the measurement and dropout processes. Given $\boldsymbol{\Omega}^{(t)}$, the current estimate for $\boldsymbol{\Omega}$, the updated estimate $\boldsymbol{\Omega}^{(t+1)}$ is obtained through one iteration of the EM algorithm, i.e., through one E step and one M step. The procedure keeps iterating between the E step and the M step until convergence is attained, i.e., until

$$\left|\ell(\boldsymbol{\Omega}^{(t+1)}|\boldsymbol{y}^o, \boldsymbol{d}) - \ell(\boldsymbol{\Omega}^{(t)}|\boldsymbol{y}^o, \boldsymbol{d})\right| < \varepsilon,$$

for some small, pre-specified $\varepsilon > 0$. More details on the EM algorithm can be found in Appendix A.

4 Classification

After fitting the latent-class mixture model to an incomplete set of repeated measurements, one could also classify the subjects examined into the different mixture components of the fitted model, i.e., into the different latent subgroups of the population. Through the structure of the latent-class mixture model, the subdivision of the population in latent groups depends on the number of observed measurements, i.e., on the dropout indicator or pattern, as well as on the values of the observed response measurements. Therefore, the classification of subjects

into different latent groups can be useful to assess the coherence between the dropout process and the measurement process. In certain cases such latent groups can have a biological or otherwise substantive meaning. For instance, subjects of one group could have higher response values and drop out earlier in the study, whereas subjects of another group have lower values but remain longer in the study.

The decision to which component of the mixture, or equivalently to which subgroup of the population, a specific subject is most likely to belong will be based on *posterior probabilities*. Recall that the group indicators Q_{ik}, for $i = 1, \ldots, N$ and $k = 1, \ldots, g$, take the value 1 if subject i belongs to group k, and 0 otherwise. We have that $P(Q_{ik} = 1) = \pi_k$, thus the component probabilities π_k, $k = 1, \ldots, g$, express how likely the ith subject is to belong to group k without taking into account either the observed response values $\boldsymbol{y}_i^o$ or the dropout indicator d_i for that subject. For this reason, the component probabilities are often called *prior* probabilities.

The *posterior* probability for subject i to belong to the kth group is given by

$$\begin{aligned} \pi_{ik} = P(Q_{ik} = 1 | \boldsymbol{y}_i^o, d_i) &= \left. \frac{f_i\left(\boldsymbol{y}_i^o, d_i | Q_{ik} = 1\right) P(Q_{ik} = 1)}{f_i(\boldsymbol{y}_i^o, d_i)} \right|_{\widehat{\boldsymbol{\Omega}}} \\ &= \left. \frac{\pi_k f_{ik}(\boldsymbol{y}_i^o, d_i | \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\alpha})}{\sum_{k=1}^{g} \pi_k f_{ik}(\boldsymbol{y}_i^o, d_i | \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\alpha})} \right|_{\widehat{\boldsymbol{\Omega}}}, \end{aligned}$$

where $\widehat{\boldsymbol{\Omega}}$ is the vector of parameter estimates resulting from the EM algorithm. This expresses how likely the ith subject is to belong to group k, taking into account the observed response $\boldsymbol{y}_i$ as well as the dropout indicator d_i of that subject. Using these posterior probabilities, we can apply the following classification rule

$$\text{Classify subject } i \text{ into component } k \quad \Longleftrightarrow \quad \pi_{ik} = \max_j \{\pi_{ij}\},$$

assigning subject i into the component to which it is most likely to belong.

However, we should be cautious with the resulting classification into latent subgroups based on the latter classification rule, since for a particular subject i, the vector of posterior prob-

abilities is given by $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{ig})$ with $\sum_{k=1}^{g} \pi_{ik} = 1$. Ideally, one of these posterior probabilities for subject i would lie close to 1, in which case the classification of this subject is obvious and likely to be correct. However, another scenario would be that two or more posterior probabilities are almost equal, of which one is the maximum of all posterior probabilities for that particular subject. For example, suppose we have $g = 2$ latent subgroups and subject i has posterior probabilities $(\pi_{i1}, \pi_{i2}) = (0.55, 0.45)$. In this case subject i would be classified into group 1 using the classification rule. However, since its probability to belong to this first group is only 10 percent more than its probability to belong the the second one, classification is not so obvious anymore and it is likely that subject i is classified into group 1, while actually it should be in group 2. We could assert this subject is in between both groups, being in a sense an outlier, or almost an 'in-lier' in the dataset. Therefore, apart from considering only the classification of subjects into the latent subgroups using the posterior probabilities, it is instructive to inspect the posterior probabilities in full. Furthermore, we can vary the number of latent groups g and explore in this way the sensitivity of the classification to the number of latent subgroups considered.

5 Simulation Study

An advantage of the latent-class mixture model is its flexible structure, which makes the model a helpful tool for analyzing incomplete longitudinal data. However, as already seen in Section 3.2, the estimation of the model parameters is based on a doubly iterative method, which we might expect to be computationally intensive. To check whether this disadvantage counterbalances the advantage of model flexibility, and to further assess performance, we conduct a simulation study. First, Section 5.1 describes a simplification of the latent-class mixture model which is used in the simulation study as well as later in the application in Section 6. The design and results of the simulation study are displayed in Sections 5.2 and 5.3, respectively.

5.1 A Simplification of the Latent-Class Mixture Model

In what follows, we will assume equal covariance matrices for the different mixture components, i.e., $D_1 = \ldots = D_g = D$, as well as equal residual covariance matrices, i.e., $\Sigma_i^{(1)} = \ldots = \Sigma_i^{(g)} = \Sigma_i$, which leads to $\boldsymbol{Y}_i | q_{ik} = 1, \boldsymbol{b}_i \sim N(X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i, \Sigma_i)$, with $\boldsymbol{b}_i \sim \sum_{k=1}^{g} \pi_k N(\boldsymbol{\mu}_k, D)$.

Further, we will simplify the general latent-class mixture model in two steps. First, it is assumed that there is only one subject-specific effect b_i, a shared intercept, which only influences the measurement process, not the dropout process. Second, the measurement process is assumed to depend on the latent variable, not in a direct way, but only through the shared intercept.

5.2 Design of the Simulation Study

We simulated 250 datasets, each containing measurements and covariate information of 100 subjects. The latent variable in the model is assumed to split the subjects into two latent subgroups with component probabilities $\pi_1 = 0.6$ and $\pi_2 = 1 - \pi_1 = 0.4$, respectively. Measurements of a continuous outcome are simulated at five time points. Further, these longitudinal data are assumed to follow a linear trend over time with intercept $\beta_0 = 9.4$ and slope $\beta_1 = 2.25$. The shared intercept follows a mixture of two normal distributions with different means for both latent groups: $\mu_1 = -4.4$ and $\mu_2 = -\frac{\pi_1\mu_1}{\pi_2} = 6.6$. In line with Section 5.1, the variances of these two normal distributions are assumed to be equal and are denoted by d^2. Finally, the second source of variation is the measurement error, with a variance σ^2.

Four different settings will be considered, based on varying these two variance parameters. In the first setting both variance parameters are chosen to be relatively small, $d = 2.0$ and $\sigma = 0.25$. While only the measurement error variance is increased in the second setting, $\sigma = 0.75$, both variance parameters are increased in the third setting, $d = 3.5$ and $\sigma = 1.00$. Up to the third setting, the chosen parameters result in a bimodal mixture distribution and consequently the simulated data of both latent groups are well separated. Since this might

improve estimation of the parameters, we consider a fourth setting with $d = 6$ and $\sigma = 2$ leading to a unimodal distribution of the data.

Finally, in the dropout model, the logistic regression is based on an intercept only, which differs for both groups, namely, $\gamma_1 = -2.5$ and $\gamma_2 = -1.25$, respectively, with corresponding probabilities 0.73 and 0.45 of completing the study.

The latent-class mixture model can now be formulated as follows. For a subject $i = 1, \ldots, 100$, belonging to latent group $k = 1, 2$, the measurement at time $j = 1, \ldots, 5$ is modelled by

$$Y_{ij} = \beta_0 + \beta_1 \text{ time}_j + b_i + \boldsymbol{\varepsilon}_{ij}^{(k)}, \tag{8}$$

with

$$b_i \sim \pi_1 N\left(\mu_1, d^2\right) + \pi_2 N\left(\mu_2, d^2\right) \quad \text{and} \quad \boldsymbol{\varepsilon}_i^{(k)} \sim N\left(\mathbf{0}, \sigma^2 \boldsymbol{I}_5\right). \tag{9}$$

Further, the dropout model is expressed as $\text{logit}[g_{ij}(\boldsymbol{w}_{ij}, \boldsymbol{b}_i, q_{ik})] = \gamma_k$.

5.3 Results of the Simulation Study

To provide insight in the nature of the four simulation settings, we randomly picked a dataset out of the 250 simulated datasets, one for each setting. The individual profiles of these datasets are shown in Figure 1. Further, Table 1 contains the results of the simulation study. Besides comparing the mean estimates and true values of the parameters through the bias, we also consider the mean squared error (MSE), simultaneously involving bias and precision.

Let us discuss the three simulation settings in turn. For the first one, Figure 1(a) shows a clear distinction between both groups, which owes to the small variance, d^2, of the mixture distribution, relative to the systematic difference between the mean of both groups, $\mu_1 - \mu_2$. Further, the small measurement error variance, σ^2, ensures the within-subject variability to be small, resulting in almost straight individual profiles. From Table 1, the mean estimates of the parameters are close to the true values, with biases of the order 10^{-2} or less. Together with small mean squared error values, of which the magnitude does not exceed 10^{-4}, this

indicates the fit of the latent-class mixture model is very close to the simulated data. This was expected due to earlier observations.

Increasing the measurement error variance in the second simulation setting leads to an increased within-subject variability. The discrepancy between both latent groups is still very obvious (Figure 1(b)). The bias increases slightly, but remains of the same order. For the MSE values, we observe a small increase, but its magnitude does not exceed 10^{-3}. So, we can conclude the model fits the data well, even with a larger within-subject variability.

In the penultimate simulation setting, not only the measurement error variance is increased, but also the variance in the mixture components. In Figure 1(c), we observe that on top of the larger within-subject variability, the gap between both latent groups now vanishes. The discrepancy between the groups seems to have vanished, and profiles appear to be homogeneous. Let us look at the results in Table 1 to see whether this has an influence on the model fit. For some of the parameters, the mean estimates deviates little from the true value. However, bias and MSE values remain small, the order of magnitude not exceeding 10^{-1} and 10^{-3}, respectively. Thus the latent-class mixture model does fit the simulated data well.

Finally, in the last simulation setting, in which even larger values for both variance parameters result in simulated data following an unimodal mixture distribution, profiles again seem to be homogeneous (Figure 1(d)). Remarkably, even in this setting, bias and MSE values remain small, both with order of magnitude below 10^{-1}.

From the four simulation settings we conclude that, whenever the model is correctly specified, it fits very well; so, this applies even when the mixture distribution is unimodal. This suggests that, for a real application, the fit is likely to be good in cases where the researcher has decent insight into the true mean structure.

Computation time increased from about 30 minutes for fitting the latent-class mixture model to a simulated dataset of the first setting, to a bit over two hours for fitting one of the

later settings. Thus, fitting the latent-class mixture model is not unreasonable in terms of computation time, perhaps against initial expectation.

6 Analysis of Depression Trial Data

We apply the latent-class mixture model to a depression trial, arising from a randomized, double-blind psychiatric clinical trial, conducted in the United States. The primary objective of this trial was to compare the efficacy of an experimental anti-depressant with placebo to support a New Drug Application. In these retrospective analyses, data from 170 patients are considered. The Hamilton Depression Rating Scale ($HAMD_{17}$) is used to measure the depression status of the patients. For each patient, a baseline assessment is available, as well as 5 post-baseline visits going from visit 4 to 8. Individual profiles of the change in $HAMD_{17}$ score from baseline for this depression trial are shown in Figure 2.

In the two subsequent sections, a latent-class mixture model is fitted to the depression trial and a sensitivity analysis performed. The latter will establish the latent-class mixture model as a viable sensitivity tool.

6.1 Formulating a Latent-Class Mixture Model

The latent-class mixture model framework is used to analyze the depression trial, assuming the patients can be split into g latent subgroups.

The mean structure is determined based on an exploratory analysis. As a result, the heterogeneity linear mixed model for the change in $HAMD_{17}$ score includes as fixed effects an intercept, the treatment variable, the baseline $HAMD_{17}$ score, the linear and quadratic time variable, and the interaction between treatment and time. In the latent-class mixture model with two group, the parameter values for these fixed effects are assumed to be equal for both latent subgroups. The measurement error terms are assumed to be independent and to follow a normal distribution with mean 0 and variance σ^2.

A shared intercept, b_i, is included in the measurement model, which follows a mixture of g normal distributions with different means, $\mu_1, \ldots, \mu_g$ respectively, but with equal variance d^2.

Further, the dropout process is modelled based on a logistic regression, including and intercept and time variable, which, in case of the two-group latent-class model, can differ between both latent subgroups ($\gamma_{0,1}, \ldots, \gamma_{0,g}$ corresponding to the intercept, and $\gamma_{1,1}, \ldots, \gamma_{1,g}$ corresponding to the slope).

At first, the same simplification as used in the simulation study in Section 5 is considered, i.e., the shared intercept is only included in the measurement model, not in the dropout model. Afterwards, we extend the model by adding the shared intercept to the dropout model as well, meaning the dropout model changes from

$$\text{logit}[g_{ij}(\boldsymbol{w}_{ij}, \boldsymbol{b}_i, q_{ik})] = \gamma_{0,k} + \gamma_{1,k}\, t_j \tag{10}$$

to

$$\text{logit}[g_{ij}(\boldsymbol{w}_{ij}, \boldsymbol{b}_i, q_{ik})] = \gamma_{0,k} + \gamma_{1,k}\, t_j + \lambda\, b_i, \tag{11}$$

where t_j is the jth visit.

An overview of the models considered is given in Table 2. Since assessing the number of components by a classical likelihood ratio test is not valid in the mixture model framework (McLachlan and Peel, 2000), we calculated the Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) for all models.

A model building exercise is performed starting with fitting a one-component latent-class mixture model, which comes down to a classical shared parameter model, as well as a two-component latent-class mixture model. Next, we compare these models using the AIC and BIC criteria, and depending on the choice made by both criteria, we decide whether we fit a latent-class mixture model with three latent subgroups.

Table 2 shows that when assuming dropout model (10), AIC opts for the model with two latent subgroups (Model 2), whereas BIC gives preference to the shared-parameter model

(Model 1). Further, in case of dropout model (11) however, both information criteria select the shared-parameter model (Model 4). Note that, since the dropout model in Model 1 does not depend on the shared intercept, the dropout model and the measurement model are independent, resulting in the MCAR assumption, whereas in Model 2, the dropout model is linked to the measurement model through the latent classes (MNAR).

Overall, the AIC criterion prefers Model 2, the 2-component latent-class mixture model with no random effect in the dropout model, whereas BIC picks Model 4, the classical shared parameter model. Since both criteria select a different model, we will take a more detailed look at the latent-class mixture model with two components, indicated by AIC, whereas we will consider the classical shared-parameter model in a sensitivity analysis in the next section.

Parameter estimates with corresponding standard errors and p-values of the two-component latent-class mixture model are shown in Table 3.

Once this latent-class mixture model has been fitted to the depression trial data, the posterior probabilities can be used to classify the patients into two subgroups as shown in Section 4. The 170 patients split into 79 and 91 patients classified into the first and second group, respectively. In Figure 3, the left panel represents the individual profiles of patients classified into the first latent group, and the right one represents the individual profiles of patients classified into the second group. Clearly, the first group corresponds to patients with lower $HAMD_{17}$ scores, that continue to decrease over time. This means these are the patients getting better. On the other hand, the second group contains patients with a higher change versus baseline compared to the patients from the first group. Their changes of $HAMD_{17}$ score fluctuate around 0, more specifically somewhere in the region between -10 and 10. In addition, without taking into account the within-subject variability, their profiles appear more or less time-constant. A more formal comparison of both latent groups regarding their change of $HAMD_{17}$ score versus baseline confirms this association between the classification and the profile over time. Furthermore, a formal test for association of baseline values and group classification is not significant, indicating similar baseline $HAMD_{17}$ scores for patients

in both groups.

Based on this difference in location of the profiles between both groups, this classification of subjects can be interpreted as being a split into acute versus chronic depression. Patients in both the acute and chronic groups enter the study with a baseline value indicating depression. However, the profiles of the patients in the acute group show recovery during the trial, whereas the depression score of patients in the chronic group remains more or less level.

Further, this difference between both latent groups is not due to treatment, since the classification of subjects in latent subgroups is independent of their treatment allocation. Indeed, the estimated odds ratio between the latent classification variable and the treatment allocation is 0.75, which was expected since the observed treatment groups are included in the mean structure of the measurement model. Moreover, when the treatment variable would be included in the dropout model, this independence would even increase.

Regarding the incompleteness of the patients in both latent groups, we notice a clear difference, which is confirmed by chi-square tests for independence, implying a significant association between the dropout pattern and the latent classification. The first latent group mainly contains patients who complete the study, 62 in total. Of the 17 patients who drop out, merely 2 drop out at visit 6, 3 more at visit 7, and 12 patients missed the last visit only. The dropout percentage in the second latent group is larger, 48.4% compared to 21.5% in the first group, or 44 out of 91 patients. Of these incompleters, 17 drop out after the first visit, 10 more at visit 6, 11 at the penultimate visit, and 6 more at the last visit.

Finally, the latent groups can also be compared by focussing on demographic characteristics such as age, gender, and origin, yielding no association between the latent classification with either gender or origin, but a significant association with age. Consequently, patients in the acute group are younger than the patients in the chronic group, with a mean age of 38.5 and 42.4, and corresponding 95% confidence intervals $[36.1, 41.0]$ and $[40.0, 44.7]$,respectively.

However, as mentioned in Section 4, using this classification rule does not render insight into

how sure the classification is in one of the two groups. This will depend on the magnitude of the maximal posterior probability. Since the latent-class mixture model considered here only contains two latent groups, we merely need to look at one of the posterior probabilities, e.g., at the posterior probability that the subject belongs to group 1, π_{i1}. Based on this π_{i1}, the subjects can be classified following the guidelines of Table 4. If the posterior probability π_{i1} lies between 0.45 and 0.55, it is uncertain to which group the subject can be classified. Only 8 out of 170 patients in the depression trial are in this situation. For most patients, 152 or 89.4%, it is clear into which group they can be classified, since their maximal posterior probability is above 0.60. Furthermore, aforementioned association of the latent classification with the location of profiles, the dropout pattern, and patient's age as well as independence of baseline values and patient's origin and gender, is confirmed by testing the independence of these variables with the posterior probabilities, which can be viewed as continuous variables ranging from 0 to 1.

6.2 A Sensitivity Analysis

In this section, we apply latent-class mixture models as a sensitivity analysis tool. In addition to the two-component latent-class mixture model shown in Section 6.1, a classical shared-parameter model will be fitted to the depression trial, as well as a pattern-mixture model, and two selection models, based on the selection models introduced by Diggle and Kenward (1994). All models contain the same fixed effects as in the two-component latent-class mixture model, i.e., intercept, treatment, time, baseline, time^2, and treatment-by-time interaction.

The classical shared-parameter model, selected by the BIC criterion in Section 6.1, includes a shared intercept $b_i \sim N(0, d^2)$, conditional upon which the measurement model follows a normal distribution $\boldsymbol{Y_i}|b_i \sim N(\boldsymbol{X}_i\beta + b_i, \sigma^2 I_{n_i})$, and the dropout process is based on (11).

Next, the Diggle-Kenward (DK) model combines a multivariate normal model for the measurement process with a logistic regression model for the dropout process. More specifically, the measurement model assumes that the vector $\boldsymbol{Y_i}$ of repeated measurements for the ith

subject satisfies the linear regression model $\boldsymbol{Y_i} \sim N(\boldsymbol{X}_i\beta, \boldsymbol{V}_i)$, $i = 1, \ldots, N$. The matrix $\boldsymbol{V}_i$ can be left unstructured or assumed of a specific form. For the depression trial, the linear mixed model (Verbeke and Molenberghs, 2000) is used to model the measurement process, with an unstructured covariance matrix. Further, let $\boldsymbol{h}_{ij} = (y_{i1}, \ldots, y_{i;j-1})$ denote the observed history of subject i up to time $t_{i,j-1}$. The DK model for the dropout process allows the conditional probability for dropout at occasion j, given that the subject was still observed at the previous occasion, to depend on the history h_{ij} and the possibly unobserved current outcome y_{ij}, but not on future outcomes y_{ik}, $k > j$. In the two models considered for the depression trial, the logistic dropout model will take the form

$$\text{logit}\left[P(D_i = j \mid D_i \geq j, \boldsymbol{h}_{ij}, y_{ij}, \boldsymbol{\Omega})\right] \;=\; \psi_0 + \psi_1 y_{i,j-1} + \psi_2 y_{ij}. \tag{12}$$

Regarding the missingness mechanism, the first selection model assumes the MAR assumption to hold, yielding $\psi_2 = 0$, whereas the second one assumes MNAR.

Finally, a pattern-mixture model is fitted by adding pattern-specific intercepts and slopes to the same multivariate normal model as used in the DK models. Notice that the classification function in the latent-class mixture model is a data driven approach to define groups, whereas pattern-mixture models use the assumptions to define groups in function of dropout patterns.

Since the main interest of the depression trial was in the treatment effect at the last visit, Table 5 shows the estimates, standard errors, and p-values for this effect under the five fitted models. Clearly, the p-values resulting from all five models are very similar and between around 0.07 and 0.11, yielding the same conclusion for the treatment effect at visit 8. Thus, the significance results are not sensitive to the model used, and hence more trust can be put into the conclusion. This is because a deflated estimate is combined with a reduced standard error. However, note that using both the two-component latent-class mixture model and the classical shared-parameter model, the standard error is reduced by 0.3 units, compared to either selection model, or pattern-mixture model, resulting in a more accurate confidence interval for the treatment effect at the last visit.

Furthermore, we explore the sensitivity of the treatment-by-time interaction by comparing

the estimates, standard errors and p-values under the five fitted models in Table 5. The p-values are clearly moving around the significance level of 0.05. Whereas under the latent-class mixture model and the shared-parameter model the p-value is about 0.03, the p-value under both selection models and the pattern-mixture model is around 0.07. While one should be cautious with over-interpretation of p-values, there are contexts, such as regulated clinical trials, where strict decision rules are implemented. In such a case and when in addition the treatment by time interaction is the primary effect, the latent-class mixture model and the shared-parameter model would lead to a claim of significance, whereas this would not be justified with neither the selection models nor the pattern-mixture model.

7 Concluding Remarks

We have proposed latent-class mixture models to analyze incomplete longitudinal data in which incompleteness is due to dropout of subjects. Through its structure, the model captures unobserved heterogeneity between latent subgroups of the population. It in an extension of the shared-parameter model, in the sense that both the measurement and dropout processes are allowed to share a set of random effects, conditional upon which both processes are assumed to be independent. It can, at the same time, be seen as an extension of the pattern-mixture model, now with latent rather than explicitly observed groups. As shown in the simulation study, the flexibility of such latent-class mixture models outweighs the expected modelling complexity.

Our proposal can be used for flexible modeling, as a sensitivity analysis instrument, and for further exploration of the latent class membership. Of course, care has to be taken when interpreting latent classes, since in some applications they may merely be artifacts, without any substantive grounds. In others, there may be more basis for their existence. We believe, together with mental health scientists, the two-component classification in our example, refers to the natural split of the patients, regardless of which treatment they were allocated to, into the more chronic and the more acute ones. An additional word of caution is needed regarding the number of latent classes to be considered. This is a tricky but well

documented problem (McLachlan and Peel 2000). A practical way out is to consider several choices for the number of components, pick the most reasonable one, and assess whether alternative choices would substantially alter the conclusions.

Evidently, the computational burden of the LCMM increases over non-latent-class models, but is still reasonable. For example, whereas the MNAR version of the Diggle and Kenward model takes around one hour and the one-component mixture needs about the same amount of time, the two-component mixture increases needs around one order of magnitude more. Furthermore, the performance of the algorithm is remarkably computationally stable, given sensible starting values (e.g., built from non-mixture classical models). Details on starting value selection are embedded in the companion manual, to be found alongside the software code on the authors' web pages.

Acknowledgements

Caroline Beunckens, Geert Molenberghs and Geert Verbeke gratefully acknowledge support from Belgian IUAP/PAI network "Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data".

References

Aitkin, M. and Rubin, D.B. (1985). Estimation and Hypothesis Testing in Finite Mixture Models. *Journal of the Royal Statistical Society, Series B*, **47**, 67–75.

Böhning, D. (1999). *Computer Assited Analysis of Mixtures (C.A.M.A.N.)*. Marcel and Decker Inc., New York.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Diggle, P.D., Kenward, M.G. (1994). Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics*, **43**, 49–93.

Follmann, D. and Wu, M. (1995). An approximate generalized linear model with random effects for informative missing data. *Biometrics*, **51**, 151–168.

Jansen, I., Beunckens, C., Molenberghs, G., Verbeke, G. and Mallinckrodt, C. (2006). Analyzing incomplete binary longitudinal clinical trial data. *Statistical Science*, **21**, 52–69.

Kenward, M.G., Molenberghs, G., Thijs, H. (2003). Pattern-mixture models with proper time dependence. *Biometrika*, **90**, 53–71.

Laird, N.M. (1994). Discussion to Diggle, P.J. and Kenward, M.G.: Informative dropout in longitudinal data analysis. *Applied Statistics*, **43**, 84.

Little, R.J.A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, **88**, 125–134.

Little, R.J.A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika*, **81**, 471–483.

Little, R.J.A. (1995). Modeling the dropout mechanism in repeated measures studies. *Journal of the American Statistial Association*, **90**, 1112–1121.

Little, R.J.A., Rubin, D.B. (1987) *Statistical Analysis with Missing Data.* New York: Wiley.

McLachlan, G.J. and Peel, D. (2000) *Finite mixture models*. New York: Wiley.

Molenberghs, G., Kenward, M.G., and Lesaffre, E. (1997). The analysis of longitudinal ordinal data with non-random dropout. *Biometrika*, **84**, 33–44.

Molenberghs, G., Michiels, B., Kenward, M.G., and Diggle, P.J. (1998). Missing data mechanisms and pattern-mixture models. *Statistica Neerlandica*, **52**, 153–161.

Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M.G., Mallinckrodt, C., and Carroll, R.J. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics*, **5**, 445–464.

Mori, M., Woodworth, G.G., and Woolson, R.F. (1992). Application of empirical Bayes inference to estimation of rate of change in the presence of informative right censoring. *Statistics in Medicine*, **11**, 621–631.

Redner, R.A. and Walker, H.F. (1984). Mixture densities, maximum likelihood, and the EM algorithm. *SIAM Review*, **26, 2**, 195–239.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.

Ten Have, T.R., Kunselman, A.R., Pulkstenis, E.P., and Landis, J.R. (1998). Mixed effects logistic regression models for longitudinal binary response data with informative drop-out. *Biometrics*, **54**, 367–383.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, **91**, 217–222.

Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.

Wu, M.C. and Bailey, K.R. (1989). Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics*, **45**, 939–955.

Wu, M.C. and Carrol, R.J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modelling the censoring process. *Biometrics*, **44**, 175–188.

A Appendix: Details on the EM Algorithm

A.1 The E Step

Let us describe the iteration step $t+1$, where the estimate is updated to $\boldsymbol{\Omega}^{(t+1)}$, using the obtained estimate from iteration step t, $\boldsymbol{\Omega}^{(t)}$. The E step consists of the calculation of the conditional expectation of $\ell(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d}, \boldsymbol{q})$, given $\boldsymbol{y}^o$ and $\boldsymbol{d}$, which is given by

$$\mathcal{O}(\boldsymbol{\Omega}|\boldsymbol{\Omega}^{(t)}) \;=\; E\left[\ell(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d}, \boldsymbol{Q}) \,\middle|\, \boldsymbol{y}^o, \boldsymbol{d}, \boldsymbol{\Omega}^{(t)}\right]$$

$$
\begin{aligned}
&= E\left[\sum_{i=1}^{N}\sum_{k=1}^{g} Q_{ik}\left\{\ln\pi_k + \ln f_{ik}(\boldsymbol{y}_i^o, d_i|\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\alpha})\right\}\middle|\boldsymbol{y}^o,\boldsymbol{d},\boldsymbol{\Omega}^{(t)}\right]\\
&= \sum_{i=1}^{N}\sum_{k=1}^{g} E\left[Q_{ik}\middle|\boldsymbol{y}^o,\boldsymbol{d},\boldsymbol{\Omega}^{(t)}\right]\left\{\ln\pi_k + \ln f_{ik}(\boldsymbol{y}_i^o, d_i|\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\alpha})\right\}.
\end{aligned}
$$

Thus, we need to calculate $E\left[Q_{ik}\middle|\boldsymbol{y}^o,\boldsymbol{d},\boldsymbol{\Omega}^{(t)}\right]$:

$$
\begin{aligned}
E\left[Q_{ik}\middle|\boldsymbol{y}^o,\boldsymbol{d},\boldsymbol{\Omega}^{(t)}\right] = P\left(Q_{ik}=1|\boldsymbol{y}^o,\boldsymbol{d},\boldsymbol{\Omega}^{(t)}\right) &= \left.\frac{f_i\left(\boldsymbol{y}_i^o, d_i|Q_{ik}=1\right)P(Q_{ik}=1)}{f_i(\boldsymbol{y}_i^o,d_i)}\right|_{\boldsymbol{\Omega}^{(t)}}\\
&= \left.\frac{\pi_k f_{ik}(\boldsymbol{y}_i^o,d_i|\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\alpha})}{\sum_{k=1}^{g}\pi_k f_{ik}(\boldsymbol{y}_i^o,d_i|\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\alpha})}\right|_{\boldsymbol{\Omega}^{(t)}} = \pi_{ik}(\boldsymbol{\Omega}^{(t)}),
\end{aligned}
$$

where $\pi_{ik}(\boldsymbol{\Omega}^{(t)})$ is the posterior probability for the ith subject to belong to the kth component of the mixture. This means the E step reduces to the calculation of posterior probabilities $\pi_{ik}(\boldsymbol{\Omega}^{(t)})$, for $i = 1,\ldots,N$ and $k = 1,\ldots,g$. Note that this also requires calculation of $f_{ik}(\boldsymbol{y}_i^o, d_i|\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\alpha})$, and consequently integration over the unknown mixture component membership to calculate, which is done numerically using Gauss-Legendre quadrature.

A.2 The M Step

The updated estimate $\boldsymbol{\Omega}^{(t+1)}$ is now obtained from maximizing $\mathcal{O}(\boldsymbol{\Omega}|\boldsymbol{\Omega}^{(t)})$ with respect to $\boldsymbol{\Omega}$. From the E step we know that $\mathcal{O}$ equals

$$
\begin{aligned}
\mathcal{O}(\boldsymbol{\Omega}|\boldsymbol{\Omega}^{(t)}) &= \sum_{i=1}^{N}\sum_{k=1}^{g}\pi_{ik}(\boldsymbol{\Omega}^{(t)})\left\{\ln\pi_k + \ln f_{ik}(\boldsymbol{y}_i^o,d_i|\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\alpha})\right\}\\
&= \underbrace{\sum_{i=1}^{N}\sum_{k=1}^{g}\pi_{ik}(\boldsymbol{\Omega}^{(t)})\ln\pi_k}_{=\mathcal{O}_1(\boldsymbol{\pi}|\boldsymbol{\Omega}^{(t)})} + \underbrace{\sum_{i=1}^{N}\sum_{k=1}^{g}\pi_{ik}(\boldsymbol{\Omega}^{(t)})\ln f_{ik}(\boldsymbol{y}_i^o,d_i|\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\alpha})}_{=\mathcal{O}_2(\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\alpha}|\boldsymbol{\Omega}^{(t)})}\\
&= \mathcal{O}_1(\boldsymbol{\pi}|\boldsymbol{\Omega}^{(t)}) + \mathcal{O}_2(\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{\Omega}^{(t)}). \qquad (13)
\end{aligned}
$$

The first term in (13) only depends on $\boldsymbol{\pi}$, whereas the second one only depends on $\boldsymbol{\theta}$, $\boldsymbol{\psi}$, and $\boldsymbol{\alpha}$. Hence, to find the maximum of the $\mathcal{O}$ function with respect to $\boldsymbol{\Omega}' = (\boldsymbol{\pi}',\boldsymbol{\theta}',\boldsymbol{\psi}',\boldsymbol{\alpha}')$, we can maximize both terms separately.

Let us first maximize the $\mathcal{O}$ function with respect to $\boldsymbol{\pi}$. This requires the maximization of $\mathcal{O}_1$, since $\mathcal{O}_2$ is independent of $\boldsymbol{\pi}$. Under the restriction $\sum_{k=1}^{g} \pi_k = 1$, we can rewrite $\mathcal{O}_1$ as follows

$$\mathcal{O}_1(\boldsymbol{\pi}|\boldsymbol{\Omega}^{(t)}) = \sum_{i=1}^{N}\sum_{k=1}^{g-1} \pi_{ik}(\boldsymbol{\Omega}^{(t)}) \ln \pi_k + \sum_{i=1}^{N} \pi_{ig}(\boldsymbol{\Omega}^{(t)}) \ln \left(1 - \sum_{k=1}^{g-1} \pi_k\right).$$

If we now set all first-order derivatives with respect to $\pi_1, \ldots, \pi_{g-1}$ equal to zero, this yields the updated estimate to satisfy

$$\begin{aligned}
\frac{\partial \mathcal{O}_1}{\partial \pi_k} = 0 \quad &\Leftrightarrow \quad \sum_{i=1}^{N} \frac{\pi_{ik}(\boldsymbol{\Omega}^{(t)})}{\pi_k^{(t+1)}} - \sum_{i=1}^{N} \frac{\pi_{ig}(\boldsymbol{\Omega}^{(t)})}{1 - \sum_{k=1}^{g-1} \pi_k^{(t+1)}} = 0 \\
&\Leftrightarrow \quad \sum_{i=1}^{N} \frac{\pi_{ik}(\boldsymbol{\Omega}^{(t)})}{\pi_k^{(t+1)}} = \sum_{i=1}^{N} \frac{\pi_{ig}(\boldsymbol{\Omega}^{(t)})}{\pi_g^{(t+1)}} \\
&\Leftrightarrow \quad \frac{\pi_k^{(t+1)}}{\pi_g^{(t+1)}} = \frac{\sum_{i=1}^{N} \pi_{ik}(\boldsymbol{\Omega}^{(t)})}{\sum_{i=1}^{N} \pi_{ig}(\boldsymbol{\Omega}^{(t)})}. \qquad (14)
\end{aligned}$$

This in turn implies that

$$\begin{aligned}
1 = \sum_{k=1}^{g} \pi_k^{(t+1)} \quad &= \quad \sum_{k=1}^{g} \frac{\pi_g^{(t+1)} \sum_{i=1}^{N} \pi_{ik}(\boldsymbol{\Omega}^{(t)})}{\sum_{i=1}^{N} \pi_{ig}(\boldsymbol{\Omega}^{(t)})} \\
&= \quad \frac{\pi_g^{(t+1)} \sum_{i=1}^{N} \overbrace{\sum_{k=1}^{g} \pi_{ik}(\boldsymbol{\Omega}^{(t)})}^{=1}}{\sum_{i=1}^{N} \pi_{ig}(\boldsymbol{\Omega}^{(t)})} = \frac{N\ \pi_g^{(t+1)}}{\sum_{i=1}^{N} \pi_{ig}(\boldsymbol{\Omega}^{(t)})},
\end{aligned}$$

and hence

$$\pi_g^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} \pi_{ig}(\boldsymbol{\Omega}^{(t)}). \qquad (15)$$

From (14) and (15) it follows that the updated estimates $\pi_k^{(t+1)}$, $k = 1, \ldots, g$, are given by

$$\pi_k^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} \pi_{ik}(\boldsymbol{\Omega}^{(t)}),$$

i.e., the updated mixture component probabilities are equal to the average posterior probabilities.

Next, to find the maximization of the $\mathcal{O}$ function with respect to $\boldsymbol{\theta}$, $\boldsymbol{\psi}$, and $\boldsymbol{\alpha}$, we need to maximize

$$\mathcal{O}_2(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\alpha}|\boldsymbol{\Omega}^{(t)}) = \sum_{i=1}^{N} \sum_{k=1}^{g} \pi_{ik}(\boldsymbol{\Omega}^{(t)}) \ln f_{ik}(\boldsymbol{y}_i^o, d_i|\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\alpha})$$

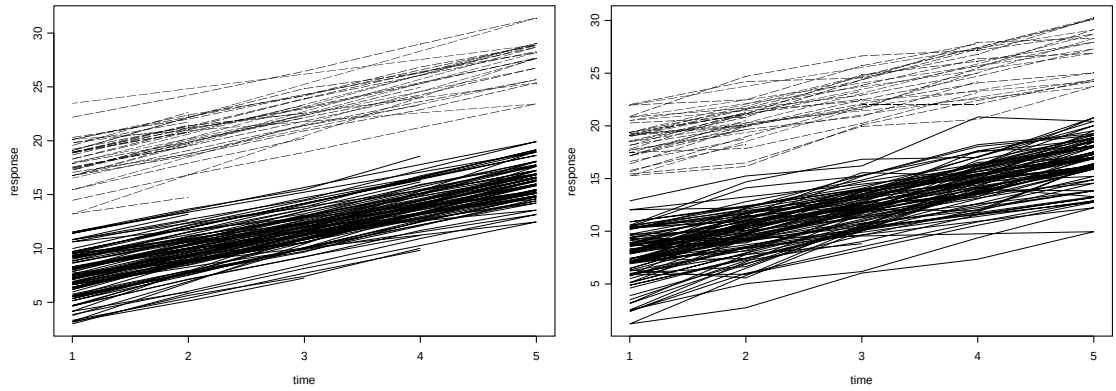
with respect to these parameters. However, in general, this cannot be done analytically. Therefore, a classical numerical maximization procedure such as, for example, Newton-Raphson is needed. Note that in such cases, the EM algorithm is doubly iterative, which might have an impact on the computation time.

A.3 Some Remarks Regarding the EM Algorithm

It can be shown that an EM step cannot decrease the likelihood value $\ell(\boldsymbol{\Omega}|\boldsymbol{y}^o, \boldsymbol{d})$, i.e.,

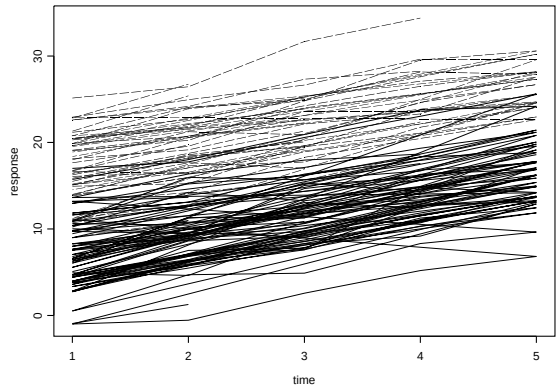
$$\ell(\boldsymbol{\Omega}^{(t+1)}|\boldsymbol{y}^o, \boldsymbol{d}) > \ell(\boldsymbol{\Omega}^{(t)}|\boldsymbol{y}^o, \boldsymbol{d}) \quad \text{for all } t.$$

This is called the monotonicity property of the EM algorithm, guaranteeing convergence of the iterative procedure, provided a finite maximum exists. However, this convergence can be painfully slow. With poorly selected starting values, such slow convergence can lead to long computation times. Apart from the local maxima resulting from the non-identifiability problem, there may be local maxima yielding different likelihood values (Böhning, 1999). This suggests that in practice multiple sets of starting values should be used. If the likelihood will have a region where it is flat, we say the likelihood has a ridge. Now, the EM algorithm is capable of converging to some particular point on that ridge, which is not the case for many other, more classical, maximization algorithms.

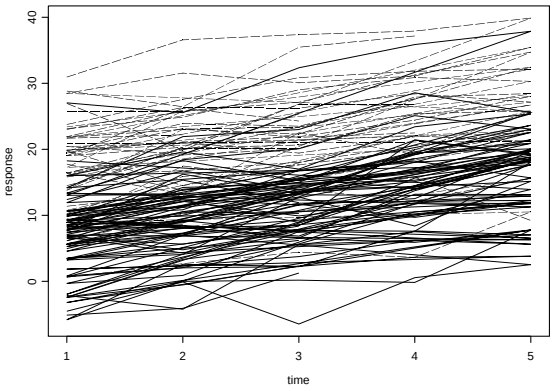


(a) Setting 1

(b) Setting 2



(c) Setting 3



(d) Setting 4

Figure 1: *Simulation Study. Individual profiles for one dataset randomly chosen out of 250 simulated datasets, for each of the three simulation settings. Dotted lines correspond to subjects from the first latent group, dashed lines to subjects from the second one.*

Table 1: *Simulation Study. Results of the simulation study: mean and true value, bias, and mean squared error (MSE) of the parameters, under the three simulations settings.*

Setting 1					**Setting 2**				
Effect	Mean	True	Bias	MSE	Effect	Mean	True	Bias	MSE
Measurement Model					*Measurement Model*				
β_0	9.37	9.40	-2.84×10^{-2}	8.07×10^{-4}	β_0	9.34	9.40	-5.75×10^{-2}	3.31×10^{-3}
β_1	2.25	2.25	1.30×10^{-4}	1.68×10^{-8}	β_1	2.25	2.25	7.56×10^{-4}	5.72×10^{-7}
σ	0.25	0.25	-2.49×10^{-4}	6.18×10^{-8}	σ	0.75	0.75	6.27×10^{-4}	3.93×10^{-7}
μ_1	-4.39	-4.40	1.31×10^{-2}	1.73×10^{-8}	μ_1	-4.36	-4.40	4.48×10^{-2}	2.00×10^{-3}
d	1.98	2.00	-1.70×10^{-2}	2.89×10^{-4}	d	1.97	2.00	-2.53×10^{-2}	6.38×10^{-4}
π_1	0.60	0.60	4.60×10^{-4}	2.12×10^{-7}	π_1	0.60	0.60	4.22×10^{-3}	1.79×10^{-5}
Dropout Model					*Dropout Model*				
γ_1	-2.52	-2.50	-2.28×10^{-2}	5.19×10^{-4}	γ_1	-2.51	-2.50	-1.26×10^{-2}	1.58×10^{-4}
γ_2	-1.26	-1.25	-1.23×10^{-2}	1.53×10^{-4}	γ_2	-1.27	-1.25	-2.30×10^{-2}	5.27×10^{-4}
Setting 3					**Setting 4**				
Effect	Mean	True	Bias	MSE	Effect	Mean	True	Bias	MSE
Measurement Model					*Measurement Model*				
β_0	9.44	9.40	3.83×10^{-2}	1.46×10^{-3}	β_0	9.59	9.40	1.92×10^{-1}	3.70×10^{-3}
β_1	2.25	2.25	1.91×10^{-4}	3.66×10^{-8}	β_1	2.24	2.25	-1.44×10^{-2}	2.06×10^{-4}
σ	0.99	1.00	-5.45×10^{-3}	2.06×10^{-5}	σ	2.01	2.00	6.07×10^{-3}	3.69×10^{-5}
μ_1	-4.69	-4.40	-2.86×10^{-1}	8.18×10^{-2}	μ_1	-4.84	-4.40	-4.39×10^{-1}	1.93×10^{-1}
d	3.43	3.50	-7.00×10^{-2}	4.90×10^{-3}	d	6.02	6.00	2.03×10^{-2}	4.10×10^{-4}
π_1	0.57	0.60	3.36×10^{-2}	1.13×10^{-3}	π_1	0.52	0.60	-8.06×10^{-2}	6.50×10^{-3}
Dropout Model					*Dropout Model*				
γ_1	-2.61	-2.50	-1.07×10^{-1}	1.14×10^{-2}	γ_1	-2.97	-2.50	-4.73×10^{-1}	2.23×10^{-1}
γ_2	-1.27	-1.25	-2.04×10^{-2}	4.17×10^{-4}	γ_2	-1.29	-1.25	-3.89×10^{-2}	1.51×10^{-3}

Table 2: *Depression Trial. Information criteria AIC and BIC, for models with dropout model (10) or (11), and* $g = 1, 2, 3$.

Model	Dropout Model	g	# Par	-2ℓ	AIC	BIC
1	$\gamma_{0,k} + \gamma_{1,k}\, t_j$	1	10	4676.07	4696.08	4727.44
2	$\gamma_{0,k} + \gamma_{1,k}\, t_j$	2	14	4662.37	**4690.37**	4734.27
3	$\gamma_{0,k} + \gamma_{1,k}\, t_j$	3	18	4662.03	4698.03	4754.48
4	$\gamma_{0,k} + \gamma_{1,k}\, t_j + \lambda\, b_i$	1	11	4669.12	4691.12	**4725.61**
5	$\gamma_{0,k} + \gamma_{1,k}\, t_j + \lambda\, b_i$	2	15	4662.02	4692.02	4739.06

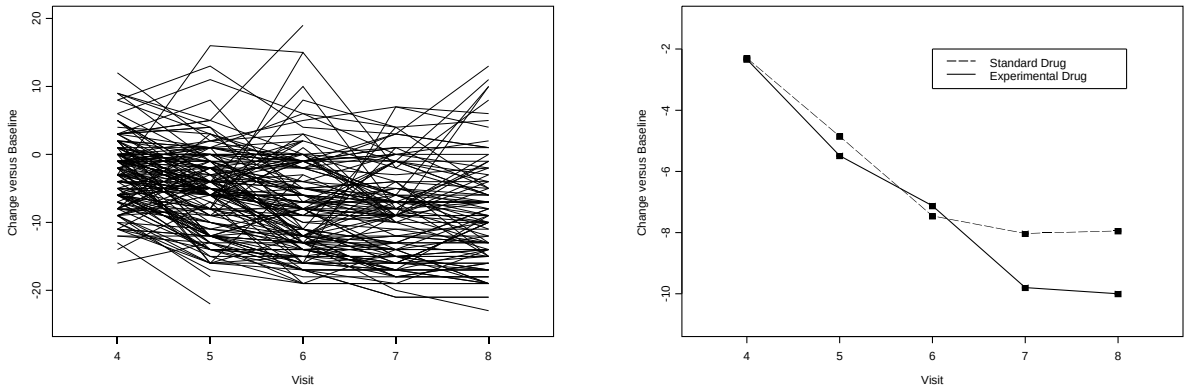


Figure 2: *Depression Trial. Individual profiles (left panel) and mean profiles by treatment arm (right panel) of the depression trial.*

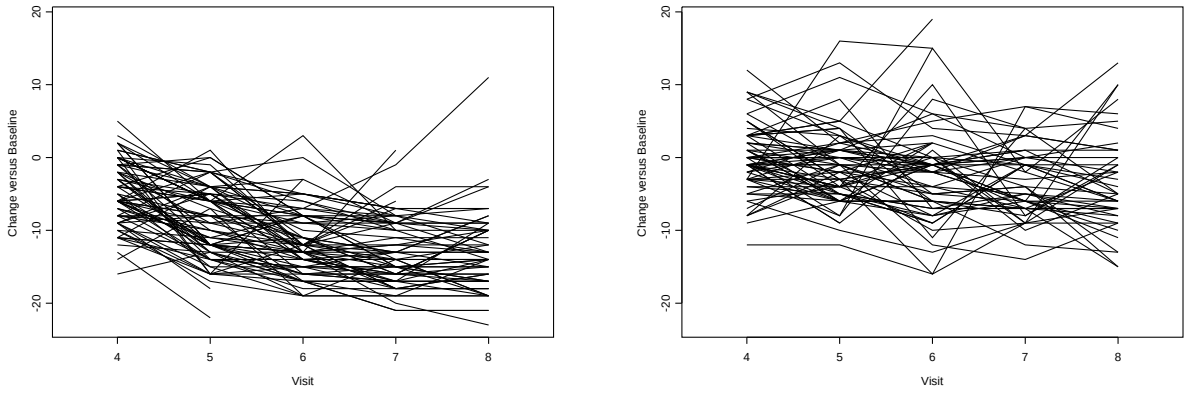


Figure 3: *Depression Trial. Classification of the subjects of the depression trial based on a latent-class mixture model. Solid lines correspond to patients classified into first group (left panel), dashed lines to patients classified into second one (right panel).*

Table 3: *Depression Trial. Parameter estimates, standard errors, and p-values for the latent-class mixture model applied to the depression trial.*

Effect	Estimate	s.e.	p-value
Measurement Model			
Intercept : β_0	23.17	3.75	< 0.0001
Treatment : β_1	2.69	1.49	0.072
Time : β_2	-6.18	1.18	< 0.0001
Time $\times$ Treatment : β_3	-0.52	0.24	0.028
Baseline : β_4	-0.42	0.07	< 0.0001
Time $\times$ Time : β_5	0.41	0.10	< 0.0001
Measurement Error : σ	4.24	0.13	< 0.0001
Dropout Model			
Intercept Group 1 : $\gamma_{0,1}$	-8.58	3.57	0.009
Time Group 1 : $\gamma_{1,1}$	0.83	0.44	0.056
Intercept Group 2 : $\gamma_{0,2}$	-1.35	1.28	0.292
Time Group 1 : $\gamma_{1,2}$	-0.05	0.20	0.793
Shared Effects			
Mean Shared Intercept Group 1 : μ_1	-3.64	0.43	< 0.0001
Variance Shared Intercept : d	2.67	0.50	< 0.0001
Prior probability Group 1 : $\pi_1 = \pi$	0.48	0.10	< 0.0001
Loglikelihood	-2331.18		

Table 4: *Depression Trial. Classification of subjects based on the magnitude of posterior probabilities* π_{i1}.

π_{i1}	Classification	# Patients
$0.80 \rightarrow 1.00$	Clearly Group 1	61
$0.60 \rightarrow 0.80$	Group 1	8
$0.55 \rightarrow 0.60$	Doubtful, more likely Group 1	5
$0.45 \rightarrow 0.55$	Uncertain	8
$0.40 \rightarrow 0.45$	Doubtful, more likely Group 2	5
$0.20 \rightarrow 0.40$	Group 2	19
$0.00 \rightarrow 0.20$	Clearly Group 2	64

Table 5: *Depression Trial. Estimates, standard errors, and p-values for the treatment effect at visit 8, as well as the treatment-by-time interaction, for the latent-class mixture model and both selection models, assuming either MAR or MNAR.*

	Treatment at Endpoint			Treatment × Time		
Model	Estimate	s.e.	p-value	Estimate	s.e.	p-value
Latent-Class Mixture Model	-1.44	0.91	0.114	-0.52	0.23	0.028
Shared-Parameter Model	-1.69	0.93	0.069	-0.50	0.24	0.035
Pattern-Mixture Model	-2.01	1.20	0.096	-0.55	0.31	0.077
MAR Selection Model	-2.17	1.25	0.082	-0.58	0.32	0.068
MNAR Selection Model	-2.16	1.24	0.081	-0.57	0.31	0.068