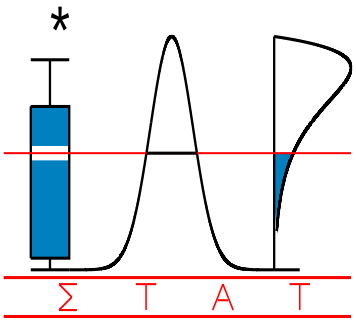
T E C H N I C A L

R E P O R T

0667

SHARED PARAMETER MODELS WITH A FLEXIBLE RANDOM EFFECTS DISTRIBUTION TO STUDY SENSITIVITY IN NON-IGNORABLE MISSINGNESS MODELS

TSONAKA, R., VERBEKE, G. and E. LESAFFRE



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

Shared Parameter Models With a Flexible Random Effects Distribution to Study Sensitivity in Non-Ignorable Missingness Models

Roula Tsonaka,[1,*] **Geert Verbeke**[1] **and Emmanuel Lesaffre**[1]

[1]Biostatistical Centre, Catholic University Leuven,

U.Z. St. Rafaël, Kapucijnenvoer 35, B–3000 Leuven, Belgium

January 15, 2007

Summary. Longitudinal studies often generate incomplete response patterns according to a Missing Not At Random mechanism. Shared parameter models provide an appealing framework for the joint modelling of the measurement and missingness processes, especially in the non-monotone missingness case, and assume a set of random effects to induce the inter-dependence. Parametric assumptions are typically made for the random effects distribution, violation of which leads to model misspecification with a potential effect in the parameter estimates and standard errors. In this paper we develop a general framework to investigate sensitivity of inferences to various distributional assumptions for the random effects within the informative missingness context. In particular, a flexible density, expressed as a mixture of normals with an unspecified mixing distribution, is assumed that captures various shapes controlled by a tuning parameter. Our proposal is illustrated on a randomized longitudinal study, with non-monotone missingness, on patients with Rheumatoid Arthritis.

Key words: Mixture model; Non-monotone missingness; Vertex-exchange method.

[*] *email:* spyridoula.tsonaka@med.kuleuven.be

1. Introduction

Missing response data often occur in longitudinal studies since subjects may be lost to follow up or do not show up on planned visits, thus resulting in monotone and non-monotone incomplete response profiles. Such missingness can be informative in the sense that the reasons for non-response depend on unobserved responses. For instance in the study that motivated this work the pain relief status of patients, suffering from Rheumatoid Arthritis, may deteriorate and hence patients decide not to appear in future scheduled visits. In this case, ignoring the missingness process can lead to biased inferences regarding the scientific questions of interest. Thus, joint modelling of the measurement and missingness processes is required to account for informative non-response.

Shared parameter models (SPMs) provide an intuitively appealing framework for the joint modelling of these two processes, in which it is assumed that a set of random effects (e.g., underlying health status) induces the inter-dependence. A key assumption underlying these models is the conditional independence (CI) of the measurement and missingness processes given the true random effects values. This CI assumption not only can be proven difficult to verify but it may also be violated when certain aspects of the model are misspecified. Two possible sources that can affect this assumption are the following. First, parametric assumptions for the shared latent process are usually made (Wu and Carroll, 1988; Wulfsohn and Tsiatis, 1997) that may not be valid since there is little information about the distribution of unobservable random effects. Second, a potential dependence of the random effects distribution on unobserved covariates will lead to model misspecification and induce heterogeneity that can cause severe bias in the fixed effects estimates (Fitzmaurice et al., 2004, pp 340). These two situations suggest that conditioning on the true random effects values can

be difficult to achieve, leading to violation of the CI assumption. This in turn implies that the probability of non-response will still depend on unobserved responses and thus certain parameter estimates and standard errors may be affected. To this end, a sensitivity analysis to varying distributional assumptions for the random effects can be invoked to investigate robustness of inferences to relaxations of common parametric assumptions. Several authors have explored methods that support the CI by either relaxing the parametric assumptions for the random effects distribution or explicitly handling population heterogeneity induced by unobserved covariates. In particular, Song et al. (2002) have expressed the shared latent component as the product of a polynomial term and the standard normal density leading thus to a smooth density that can assume several shapes. Lin et al. (2002), Roy (2003) and Lin et al. (2004) on the other hand, have explored a potential subpopulation structure by means of a latent class model. In these approaches, a class indicator with a pre-specified support size is assumed to induce the association between the longitudinal and event process, thereby allowing a distinct behavior within each group for both processes. However, none of these approaches focuses on investigating the impact of varying distributional assumptions on the resulting inferences.

In this paper, we propose a method to perform sensitivity analysis within the SPM that assumes flexible shapes for the random effects distribution controlled by a tuning parameter. In particular, we assume that the association between the involved processes is induced by a discrete latent variable G with unknown support size. This variable is incorporated into the random effects component leading to a random effects density that is expressed as a mixture with an unspecified number of normal components. Tuning the variance of the mixture normal components, various shapes for the random effects density, supported by the data, can be captured. In addition,

the use of a mixture model provides a natural framework to explore a potential population heterogeneity, since G acts as a latent group indicator. Parameter estimation for the SPM is based on a semi-parametric method, since G has an unspecified support size, and is implemented by means of the Vertex-Exchange algorithm (Böhning, 1985; Böhning, 1999). This is a gradient-based estimation method, particularly suited for semi-parametric mixture models, that is ensured to converge globally contrary to the EM algorithm. The proposed model formulation resembles the random effects distribution considered by Magder and Zeger (1996) in the ignorable linear mixed model framework in which the EM algorithm was used as estimation procedure.

Our method is exemplified on the primary endpoint of a randomized study designed to evaluate the efficacy of the trial drug in patients with Rheumatoid Arthritis (Furst et al., 2002). Rheumatoid Arthritis is a chronic inflammatory disease that causes pain, swelling and stiffness in multiple joints. In our study, 895 patients have been randomized to 5 treatment groups. Their disease activity was evaluated on a 10cm Visual Analogue Scale (VAS) and recorded on 5 planned visits. However, the disease activity can have a direct impact on the patients' quality of life and on their visiting behavior. As a result, 47% of the patients missed at least one visit in a non-monotone pattern. Reasons for a missed visit were lack of efficacy, unexpected worsening of the disease, and others. This implies a considerable loss of information (19% missingness) that could affect inference unless properly handled. The proposed sensitivity analysis framework is considered to (1) investigate the robustness of inference to various distributional assumptions for the shared latent component and (2) uncover a potential subpopulation structure in the data.

The paper is organized as follows: Section 2 presents the proposed modelling framework, Section 3 summarizes some theoretical results, Section 4 gives the details

for the estimation procedure, Section 5 evaluates the performance of the proposed model via simulation, Section 6 presents the results of the analysis of the clinical trial dataset and finally in Section 7 we summarize and conclude.

2. The Semi-Parametric Shared Parameter Model

2.1 *The Shared Parameter Model*

Suppose that for a random sample of size n, Y_i, with elements y_{ij} $(i = 1, \ldots, n;\ j = 1, \ldots, n_i;\ \max(n_i) = \mathcal{N})$, denotes the vector of the observed longitudinal responses for the ith individual, and R_i is the corresponding sequence of response indicators $r_{ij} \in \{0, 1\}$, with 1 denoting that y_{ij} is observed and 0 otherwise. Thus, here we mainly focus on the more general non-monotone case rather than on the most common dropout case.

Under the SPM framework, the joint density of the Y and R processes for the ith individual is factorized as

$$f(Y_i, R_i;\ \theta) = \int f(Y_i \mid b_i;\ \theta)\ f(R_i \mid b_i;\ \theta)\ f(b_i;\ \theta)\ db_i, \tag{1}$$

where $f(Y_i \mid b_i;\ \theta)$ and $f(R_i \mid b_i;\ \theta)$ are the conditional, on the random effects values b_i, densities for the Y and R processes respectively, $f(b_i;\ \theta)$ is the random effects density and θ the parameter vector. The model formulation (1) posits that the two processes are linked at the individual level through the random effects, and assumes that conditioning on the true random effects values, an ignorable mechanism applies, in which the non-response probability is independent of unobserved responses.

For the longitudinal responses Y_i we assume the linear mixed effects model (LMM)

$$Y_i = X_i\beta^{(1)} + Z_i(\beta^{(2)} + b_i) + \varepsilon_i, \tag{2}$$

where $\beta^{(1)}$ and $\beta^{(2)}$ denote regression coefficient vectors, X_i is the design matrix for fixed effects with no random effects counterpart, Z_i is the design matrix for covariates

with both fixed and random effects and b_i is the q-dimensional random effects. The error terms ε_i are assumed independent of b_i, and follow a multivariate mean zero normal distribution with variance σ^2 and an unstructured correlation matrix $\mathcal{P}_i(\rho)$ that is parameterized by the vector ρ and depends on i through its dimension only, i.e., $\varepsilon_i \sim N_{n_i}(0, \sigma^2\mathcal{P}_i(\rho))$. For the missingness process R, the probability of response, $p_{ij} = Pr(r_{ij} = 1)$, is modelled using a mixed effects logistic regression

$$\text{logit}(p_{ij}) = w_{ij}^T\alpha^{(1)} + z_{ij}^T(\alpha^{(2)} + \gamma\, b_i),$$

where w_{ij} is the jth row of the fixed effects design matrix W_i, $\alpha^{(1)}$ and $\alpha^{(2)}$ the regression coefficient vectors, z_{ij} the jth row of Z_i and the superscript T denotes the transpose matrix. As above covariates in Z_i are not included in W_i. The measurement and missingness processes are linked through the random effects term and their association is quantified by the parameter γ.

2.2 *The Semi-Parametric Model*

In this paper we relax the common parametric assumptions for the random effects distribution and assume that it is a mixture with mixing distribution G. In particular, the marginal density of the random effects b_i is given by

$$f(b_i; G) = \int_{\Omega_{\mathcal{M}}} f(b_i \mid \mu, D)\, dG(\mu), \tag{3}$$

where $f(b_i \mid \mu, D)$ is a kernel density with location parameter $\mu \in \mathcal{M}$, where $\mathcal{M}$ denotes the parameter space of μ, and a potential nuisance parameter D. Here we have chosen a q-variate normal density kernel with mean vector μ and covariance matrix D. Equation (3) postulates that μ is distributed according to the distribution G. In general, G can be a member of the set $\Omega_{\mathcal{M}}$ of all distribution functions, including both discrete and continuous. However, Laird (1978) and Lindsay (1983)

have shown that the non-parametric maximum likelihood estimate (NPMLE) of G, is discrete with finite support, and thus $\Omega_{\mathcal{M}}$ reduces to the set of discrete distributions. Incorporating this result in (3) leads to a mixture density with an unspecified number of normal components, i.e.,

$$f(Y_i, R_i; G, \theta) = \int f(Y_i \mid b_i; \theta_Y)\ f(R_i \mid b_i; \theta_R)\ \sum_c \pi_c\ f(b_i \mid \mu_c, D)\ db_i, \tag{4}$$

where $\theta^T = (\theta_Y^T, \theta_R^T)$ is the parameter vector for the Y and R processes, μ_c are the support points and π_c the corresponding weights of G. We call the model defined by (4) Semi-Parametric SPM (SPSP), since a parametric assumption is made for each mixture component whereas the number of components is determined from the available data. In this case, the parameter vector (θ, D, μ_c, π_c) is high dimensional and often interest lies in the parameter vector θ, whereas μ_c and π_c are treated as nuisance parameters. The covariance matrix D controls the flexibility of the random effects density allowing thereby for various shapes. In particular, as it has also been noted in Magder and Zeger (1996), when $\det\{D\} \to 0$, where $\det\{.\}$ denotes the determinant, the estimated density approaches the NPMLE, whereas large $\det\{D\}$ values lead to an oversmoothed density. Intermediate $\det\{D\}$ values would reveal potential skewness, kurtosis and multimodality in the random effects density.

The proposed formulation (4) enjoys a number of appealing features. First, the common normality assumption for the random effects component is relaxed by assuming additional variability in its mean parameter introduced by G. In addition, since no particular assumption is made on the distribution of the shared latent process, flexible shapes and association structures between the involved processes can be achieved controlled by the component covariance matrix D. This suggests that our proposal provides a general framework for sensitivity analysis under the Missing Not

At Random mechanism since the shape of the random effects distribution implicitly determines the dependence of the missingness process on the missing response components. To see this, consider the complete response vector Y_i that is decomposed into Y_i^o and Y_i^m, the observed and missing components, respectively. Then the conditional distribution of R_i given the complete data Y_i^o, Y_i^m is given by

$$f(R_i \mid Y_i^o, Y_i^m) \;\; = \;\; \int f(R_i \mid b_i) f(b_i \mid Y_i^o, Y_i^m)\, db_i,$$

which implies that the probability of non-response depends on Y_i^m only through the posterior $f(b_i \mid Y_i^o, Y_i^m)$. The choice for the prior $f(b_i)$ affects the shape of the posterior of b_i, especially when few measurements per individuals are observed. Finally, the proposed model provides a natural framework for exploring a potential subpopulation structure and classify the individuals based on their posterior probabilities. In particular, this probability for the ith individual to belong to the cth group is $f(\mu_c \mid Y_i, R_i;\ \hat{\theta}, \hat{G})$, where $\hat{\theta}$ and $\hat{G}$ are the model based estimates. The individuals are classified according to their maximum posterior probability.

3. Estimation Procedure - Geometric Results

Maximization of the log-likelihood function $\ell_n(\theta, G)$ corresponding to model (4) requires the search for the optimal set $(\hat{\theta}, \hat{G}) = (\hat{\theta}, \hat{\mu}_1, \ldots, \hat{\pi}_1, \ldots)$ that satisfies $\ell_n(\hat{\theta}, \hat{G}) = \sup_{\theta \in \Theta, G \in \Omega_{\mathcal{M}}} \ell_n(\theta, G)$. In the LMM context, the EM algorithm has been used to obtain the NPMLE of the mixing distribution G (Laird, 1978; Magder and Zeger, 1996). Since the EM algorithm requires prespecification of the support size, two approaches are in use. First, the model is fitted with a relatively large number of mixture components. This approach often suffers from numerical instability since many components will have zero weight which lies on the boundary of the parameter space. Second, the model is repeatedly fitted with an increasing number of components as long as

distinct support points with positive weight are returned. In this case, the computational burden will become considerable since many model fits would be needed for a moderate true number of components.

We consider here the Vertex-Exchange Method (VEM), suggested by Böhning (1985), to estimate G. This is a stable algorithm that can reliably maximize flat likelihood surfaces encountered in mixture models. This algorithm iteratively maximizes the log-likelihood $\ell_n(G)$ in the set $\Omega_{\mathcal{M}}$ of all discrete distributions over a prespecified grid $\mu_1, \ldots, \mu_C$ (see Section 3.1), for large C, based on the properties of the directional derivative of the log-likelihood from one distribution G_1 to another G_2. In particular, in each iteration the algorithm searches for the direction that maximizes the log-likelihood increase and weights are exchanged between points that contribute the least and the most. An approximate NPMLE of G is obtained when a very dense grid is chosen.

3.1 *Estimation of G*

The search for the optimal $\hat{G}$ is not straightforward and thus some theoretical results will be briefly presented. Lindsay (1983) has shown that, for every θ, the log-likelihood

$$\ell_n(G) = \sum_{k=1}^{K} \delta_k \log f(Y_k, R_k; G, \theta), k = 1, \ldots, K, \tag{5}$$

where δ_k is the multiplicity of $f(Y_k, R_k; G, \theta)$ with $\sum_{k=1}^{K} \delta_k = n$, is a concave functional over the set of all convex combinations of $f(Y_k, R_k; G, \theta)$. Thus, provided that the log-likelihood is bounded, there is a unique $\hat{f} = (\hat{f}(Y_1, R_1; G, \theta), \ldots, \hat{f}(Y_K, R_K; G, \theta))$ that maximizes $\ell_n(G)$. The corresponding maximum likelihood (ML) estimate $\hat{G}$, known as the NPMLE of G (Laird, 1978), is discrete with at most K support points. However, a common problem in the mixture setting is that there is no unique combination of μ_c

and π_c that maximizes $\ell_n(G)$. In other words, $\hat{G}$ is not unique, but fixing its support points μ_c will lead to a unique solution. In this case, for reasons explained in Section 4.1, characterizations of $\hat{G}$ can be achieved in terms of directional derivatives and algorithms for finding $\hat{G}$ are constructed based on the general Mixture ML Theorem (Lindsay, 1983; Simar, 1976; Jewell, 1982). The directional derivative $\mathcal{D}(G_1, G_2)$ of $\ell_n(G)$ at G_1 in the direction of G_2 is defined as

$$\mathcal{D}(G_1, G_2) = \lim_{s \to 0} \frac{\ell_n((1-s)G_1 + sG_2) - \ell_n(G_1)}{s} = \sum_{i=1}^{n} \frac{f(Y_i, R_i; G_2, \theta) - f(Y_i, R_i; G_1, \theta)}{f(Y_i, R_i; G_1, \theta)}.$$

When G_2 is degenerate at μ, then $G_2 = G_\mu$ and the directional derivative gets $\mathcal{D}(G_1, G_\mu) = \sum_{i=1}^{n} f(Y_i, R_i; \mu, \theta)/f(Y_i, R_i; G_1, \theta) - n$. Alternatively, characterizations can be achieved using the gradient function of $\ell_n(G)$

$$d(\mu, G_1) = \frac{1}{n} \sum_{i=1}^{n} \frac{f(Y_i, R_i; \mu, \theta)}{f(Y_i, R_i; G_1, \theta)}, \tag{6}$$

which is related to $\mathcal{D}(G_1, G_\mu)$ through

$$\mathcal{D}(G_1, G_\mu) = nd(\mu, G_1) - 1. \tag{7}$$

For the estimation of G the VEM, described in Section 4.1, is used that is based on the following conditions of the general ML theorem:

1. $\hat{G}$ maximizes $\ell_n(G)$ if and only if, for all $\mu \in \mathcal{M}$, $\mathcal{D}(\hat{G}, G_\mu) \leq 0$ or $d(\mu, \hat{G}) \leq 1$.

2. μ is in the support of $\hat{G}$ only if $\mathcal{D}(\hat{G}, G_\mu) = 0$ or $d(\mu, \hat{G}) = 1$.

4. Optimization Procedure

4.1 *Algorithm*

An iterative two-step procedure is proposed for obtaining the optimal set $(\theta, \pi_1, \ldots, \pi_C)$ that maximizes the observed data log-likelihood $\ell_n(\theta, G)$. In the first step, for fixed θ, $\ell_n(G)$ is maximized in the set $\Omega_{\mathcal{M}}$ of all discrete distributions G over a prespecified

grid $\mu_1, \ldots, \mu_C$ by means of the VEM algorithm. In the second step, for the estimated G of the first step, θ is estimated using quasi-Newton.

Initial values: Initial values $\theta^0 = (\theta_Y^0, \theta_R^0)$ for the parameters of the Y and R processes, can be obtained by fitting the appropriate ignorable mixed effects models, whereas an initial G^0 may be defined by setting the weights π_c^0 equal to $1/C$.

VEM step: Estimate G^1 given θ^0, which implies that estimation of the weights π_c^1 is only required. As it has been explained in Section 3.1, for fixed θ the problem reduces to one of maximizing a concave functional over a convex set. The VEM is a gradient based algorithm that is built on the idea of searching at each iteration the direction, among a prespecified grid of support points, that increases the likelihood. More precisely, the VEM at each iteration exchanges weights between the points that contribute the least and the most to the likelihood increase (denoted respectively as μ^- and μ^+ with corresponding weights π^- and π^+). In particular, convex combinations of the form $G^0 + s\pi_{\mu^-}\{G_{\mu^+} - G_{\mu^-}\}$ are considered where $s \in [0, 1]$ is the step length, G^0 denotes the current approximation to G and G_{μ^+} and G_{μ^-} are degenerate distributions on the points μ^- and μ^+. A first order approximation of $\Delta = \ell(G_0 + s\pi_{\mu^-}\{G_{\mu^+} - G_{\mu^-}\}) - \ell(G_0)$ around $s = 0$ results in $s\pi_{\mu^-}\{\mathcal{D}(G_0, G_{\mu^+}) - \mathcal{D}(G_0, G_{\mu^-})\}$ and implies that maximization of Δ is equivalent to identifying the points μ^- and μ^+ that respectively minimize and maximize $\mathcal{D}(G_0, G_\mu)$ over μ. In addition, based on (7) the gradient function (6) can equivalently be used, to identify μ^- and μ^+ which is easier to compute than $\mathcal{D}(.)$. Once these points have been identified, their weights are exchanged according to $\pi_{\mu^-}^1 = (1 - s)\ \pi_{\mu^-}^0$ and $\pi_{\mu^+}^1 = s\ \pi_{\mu^-}^0 + \pi_{\mu^+}^0$, in which an update for the step length s is found as the maximizer of $\ell(G^1) - \ell(G^0)$, using a line search method.

quasi-Newton step: θ^1 is estimated for fixed G^1. Since there are no closed form solutions numerical procedures are used. Here we use a quasi-Newton algorithm which requires the derivatives of (5) with respect to θ, that are presented in the Appendix.

Repeat 2 and 3 until convergence. The algorithm has converged when both conditions have been satisfied: (1) $\max_\mu d(\hat{G}^{(it)}, \mu) < 1 + \epsilon$, which guarantees that $\ell(\hat{G}^{(it)} \mid \hat{\theta}^{(it-1)}) - \ell(\hat{G}^{(it-1)} \mid \hat{\theta}^{(it-1)}) < \epsilon$, and (2) $\ell(\hat{G}^{(it)}, \hat{\theta}^{(it)}) - \ell(\hat{G}^{(it-1)}, \hat{\theta}^{(it-1)}) < \epsilon'$ for small ϵ, ϵ' (i.e., 10^{-3} and 10^{-8}, respectively) with it denoting the iteration.

4.2 *Implementation Issues*

For the implementation of the two-step procedure, a number of issues must be considered. First, the integral over the random effects in (4) is approximated using the Gauss-Hermite quadrature rule. Second, according to Section 3.1, obtaining a unique $\hat{G}$ requires fixing its support points μ_c. Since $\hat{G}$ has a finite number of support points, a finite grid with points $\mu_1, \ldots, \mu_C$ is specified and thus the choice of an appropriate range and choice of the number of points is required. Regarding the range, a grid is chosen for the scaled random effect $b_i^* = S_b^{-1} b_i$ instead of b_i to avoid specification of an unnecessarily huge grid. The matrix S_b represents an approximate estimate of the choleski decomposition of the random effects covariance matrix and is chosen equal to the choleski decomposition of $\hat{D}_b^Y$ of the corresponding ignorable measurement model. Thus, a grid for b_i^* defined in $[-\nu, \nu]^q = [-\nu, \nu] \times \ldots \times [-\nu, \nu]$ with $\nu = 3$ or 4 would in most cases be sufficient. Regarding the number of grid points, a dense grid must be chosen such that the maximum distance between the true unknown mixture means $\mu_{true} \in [\mu_{c-1}, \mu_c]$ and the assumed grid points μ_c is sufficiently small, i.e., $(\mu_c - \mu_{c-1})/2 \leq 0.1$. Thereby the resulting solution will not differ considerably from

the NPMLE. The effect of the number of grid points considered is further investigated in Section 6. We should note that the computing time required to perform a single iteration of the algorithm increases directly with the number of grid points.

A feature of the mixture density (4) that we have not discussed so far is that, in general, $E(b_i^*) = \sum_{c=1}^{C} \pi_c \mu_c \neq 0$ contrary to the commonly used zero mean normal distribution. This may lead to an overparameterized and unidentified model. To circumvent this we fix, through the iterative procedure, the parameters $\beta^{(2)}$ in model (2) at the value estimated by the ignorable mixed effects model. At the end of the estimation procedure, $\beta^{(2)}$ will be updated by $\beta^{(2)} + \hat{S}_b \sum_{c=1}^{C} \hat{\pi}_c \mu_c$.

Finally, the covariance matrix D controls the flexibility of the random effects density that can vary from an almost discrete to an oversmoothed solution. As a practical guideline, choosing values of $\det\{D\}$ in $(0, \det\{d\ \hat{D}_b^Y\}]$, where $\hat{D}_b^Y$ is the estimated by the ignorable LMM random effects covariance matrix and $d > 0$, would in most cases capture all these shapes. According to our experience, for the univariate case $d = 1/4$ is sufficient.

4.3 *Properties of the MLEs*

Assuming that the log-likelihood $\ell_n(\theta, G)$ satisfies the regularity conditions specified by Kiefer and Wolfowitz (1956) both $\hat{G}$ and $\hat{\theta}$ are consistent. In addition, the maximum likelihood estimator for θ is asymptotically normal with covariance equal to the inverse of the efficient information matrix (see van der Vaart, 1996). However, in Section 6 approximate standard errors are estimated using the second derivative matrix of the profile log-likelihood $\ell_n(\theta \mid \hat{G})$ evaluated at $\hat{\theta}$. To evaluate the quality of the standard error estimates under $\ell_n(\theta \mid \hat{G})$ we use the non parametric Bootstrap method and a close similarity has been observed. This result has also been corrob-

orated by Follmann and Lambert (1989) in which the random effects distribution in mixed effects logistic regression is estimated using non-parametric ML.

5. Simulation Study

The aim of this simulation study is first to evaluate the performance of the proposed model, and second to investigate the sensitivity of inference to violations of the commonly used normality assumption. Thus, comparisons between the SPSP presented in Section 2.2 and the SP (1) that assumes normal random effects will be made.

5.1 *Set-up*

The longitudinal process Y is simulated from a LMM with linear predictor: $\eta_{ij}^{Y} = \beta_0 + \beta_1 t_{ij} + \beta_2 \mathcal{T}_i t_{ij} + b_i + \epsilon_{ij}$, where the subscripts $i = 1, \ldots, n$ and $j = 1, \ldots, \mathcal{N}$ denote the subject and visit, respectively, t_{ij} is the time variable, $\mathcal{T}_i$ the treatment indicator and b_i the random effects component. The parameter vector is taken $(\beta_0, \beta_1, \beta_2) = (-1.2, 0.5, -1.5)$. Note that $\mathcal{T}_i$ is not included in η_{ij}^{Y} as a main effect because a randomized study is considered. For the error component, we assume $\varepsilon_i \sim N_{\mathcal{N}}(0, \sigma_Y^2)$ with $\sigma_Y = 0.5$. The sample size was chosen $n = 50, 200$ and $\mathcal{N} = 5, 15$ equally spaced visit times were assumed. Regarding the R process, a model that allows for non-monotone missingness was considered. That is the binary indicator r_{ij} was simulated from a mixed effects logistic regression with linear predictor $\eta_{ij}^{R} = \alpha_0 + \alpha_1 \mathcal{T}_i + \alpha_2 t_{ij} + \gamma b_i$, where $(\alpha_0, \alpha_1, \alpha_2) = (1.6, 2.5, -0.5)$. The assumed values for the regression parameters have been chosen such that the percentage of missingness is around 30% when $\mathcal{N} = 5$ and 60% when $\mathcal{N} = 15$. The Y and R processes are linked through the shared random intercepts b_i, and their association is measured by the parameter $\gamma = 0.7$. According to the parameterization used in Section 3, $\beta^{(1)} = (\beta_1, \beta_2)$, $\beta^{(2)} = \beta_0$, $\alpha^{(1)} = (\alpha_1, \alpha_2)$ and $\alpha^{(2)} = \alpha_0$. Regarding the

random effects component b_i four cases are considered: (i) a unimodal distribution, i.e., $N(0, \sigma_b^2)$ with $\sigma_b = 2$, (ii) a bimodal symmetric mixture of normal components, i.e., $0.5 \times N(-2, 0.5^2) + 0.5 \times N(2, 0.5^2)$, (iii) a bimodal symmetric mixture of Student's t components with 3 df, i.e., $0.5 \times t_3(-2, 0.5^2) + 0.5 \times t_3(2, 0.5^2)$ and (iv) a skewed distribution, i.e., Log-Normal$(0, 0.97^2)$. The parameter values for the assumed random effects distribution have been chosen such that for all scenarios $Var(b_i) \simeq 4$. Combining the above choices results in 16 scenarios, for which 500 datasets were simulated. Each dataset was fitted under 2 different models, namely the SPSP and the SP model. Comparisons between parameter estimates are based on the Root Mean Squared Error (RMSE) over the 500 datasets. Comparisons of the fitted random effects densities are made via the Mean Root Integrated Square Error (MRISE).

5.2 *Results*

The results of the simulation study are shown in Table 1 which contains the true values and the estimates of the model parameters under the different settings. In addition, among the various scenarios we present for each distributional assumption only the case with $n = 50$ and $\mathcal{N} = 5$. For the other scenarios the main conclusions are discussed.

[Table 1 about here.]

When the true random effects distribution is normal, both the SPSP and SP produce estimates with very close RMSE values for all parameters and even in the small sample case. However, the "correct" SP provides a better approximation to the true normal density than the SPSP. This is due to the fact that the one estimated by the SPSP model density is less smooth. When the longitudinal measurements are increased to $\mathcal{N} = 15$ or when $n = 200$ the same conclusions can be drawn.

As expected, increasing the sample size improves the efficiency and more accurate estimates are obtained. Thus, the simulation results under the unimodal scenario suggest that our proposed method can be successfully used even if heterogeneity is not present. Besides, such information is usually not available in practice.

When the random effects are not normally distributed, the SPSP model produces more efficient and less biased parameter estimates in comparison to the SP model. The results are more pronounced for the skewed scenario, i.e., the log-Normal distribution. In addition, the SPSP model gives more accurate estimates of the true density than the SP according to the values of the RMSE in all non-normal scenarios considered.

6. Application

A double-blind randomized clinical trial with 895 patients suffering from Rheumatoid Arthritis has been recently conducted in order to evaluate the safety and efficacy of the trial drug. The patients were randomized to 5 treatment groups i.e., 3 doses of the trial drug, standard treatment and placebo, and they were followed for 3 months after randomization. Follow-up was scheduled at Day 0, Week 2, 4, 8 and 12, and the endpoint we consider here is the evaluation of the disease activity by the patient on a 10cm VAS (0cm for no disease activity and 10cm for high disease activity). However, for various reasons some patients missed some of their scheduled visits resulting in non-monotone missingness that reaches 19%. The percentage of missingness per visit is 5%, 18%, 28%, 38% and 7%, respectively. The low percentage at the last visit is explained by the protocol condition that all patients should be reached at the end of the trial by any means (e.g., by phone), even if they did not show up at the trial center for their last visit. Thus it is reasonable to assume that the reason for non-response at the last visit is not related to the VAS score (Missing Completely At

Random (Rubin, 1976)), since missingness is due to contact failure. In the sample there are 19 observed patterns of missingness with the completers representing almost 53%. In order to model the evolution of the disease activity while adjusting for non-monotone missingness we use the proposed SPSP framework. Thereby, we investigate the effect of various distributional assumptions for the random effects on the resulting inferences. In addition, a potential grouping of the study population is explored and identification of the profiles of the non-compliant individuals is made.

The longitudinal evaluations of the disease are modelled using a LMM with linear predictor $\eta_{ij}^{Y} = x_{ij}^{T}\ \beta + b_i$, where $i = 1, \ldots, 895$, $j = 1, \ldots, n_i$, β denotes the vector of the regression parameters, $x_{ij} = [1, (1, \mathcal{T}_{2i}, \mathcal{T}_{3i}, \mathcal{T}_{4i}, \mathcal{T}_{5i}) \times t_{ij}, (1, \mathcal{T}_{2i}, \mathcal{T}_{3i}, \mathcal{T}_{4i}, \mathcal{T}_{5i}) \times t_{ij}^2, \text{Gender}_i, \text{Age}_i, \text{BMI}_i, \text{DaysD}_i]$ denotes the jth row of the fixed effects design matrix, b_i is the subject-specific random intercept and n_i is the number of measurements for the ith patient, with $\max(n_i) = 5$. The covariate BMI denotes the Body Mass Index and DaysD is the recorded number of days that the patient was provided with treatment. The treatment effect $\mathcal{T}_{2i}, \ldots, \mathcal{T}_{5i}$ was not included as a main effect in η_{ij}^{Y} due to randomization and since at the first visit, i.e., Day 0, there was no drug administration. In the explanatory analysis, the longitudinal outcome was fitted under the ignorable LMM, and revealed that a serial correlation structure in addition to the random intercepts component is required. Thus, in the SPSP model it is also assumed that $\varepsilon_i \sim N_{n_i}(0, \sigma^2 \mathcal{P}_i)$, where the matrix $\mathcal{P}_i$ describes the correlation of the within-subject errors ε_i. In particular, a general correlation structure is used, with the correlation between different time points represented by a different parameter i.e., $cor(\varepsilon_{ij}, \varepsilon_{ij\prime}) = \rho_{jj\prime}$, for $j \neq j\prime$.

The missingness profiles are modelled using a mixed effects logistic regression with linear predictor $\eta_{ij}^{R} = w_{ij}^{T}\ \alpha + \gamma\ b_i$, where $i = 1, \ldots, 895$, $j = 1, \ldots, 4$, α

is the regression parameter vector and $w_{ij} = [1, \mathcal{T}_{2i}, \mathcal{T}_{3i}, \mathcal{T}_{4i}, \mathcal{T}_{5i}, (1, \mathcal{T}_{2i}, \mathcal{T}_{3i}, \mathcal{T}_{4i}, \mathcal{T}_{5i}) \times t_{ij}, (1, \mathcal{T}_{2i}, \mathcal{T}_{3i}, \mathcal{T}_{4i}, \mathcal{T}_{5i}) \times t_{ij}^2]$ denotes the jth row of the fixed effects design matrix. The Y and R processes are linked through the random effect b_i and their association is measured by the parameter γ. The probability of response at the jth visit was modelled only for the first 4 visits, since only for these visits a non ignorable missingness mechanism may occur, whereas for the last visit missingness is assumed completely at random and thus the probability of response was fixed at 1.

According to Section 2.2, b_i is modelled as a mixture density with an unspecified number of normal mixture components. The model is estimated according to the procedure described in Section 4.1. Initial values for the model parameters of the Y and R process are obtained, respectively, from the ignorable LMM and the mixed effects logistic regression with normal random effects. In order to investigate the robustness of the algorithm to starting values, multiple sets of initial values have been considered that varied around the ignorable models estimates. The grid of support points for b_i^* was chosen in $[-3, 3]$ that corresponds to $[-3\hat{\sigma}_b^Y, 3\hat{\sigma}_b^Y] = [-6, 6]$ for b_i, where $\hat{\sigma}_b^Y$ is the ignorable LMM estimate of the random effects standard deviation. Wider grids have been also considered at the expense of computing time but the resulting solutions did not change. The involved integral over the random effect is approximated using Gauss-Hermite quadrature rule with $h = 51$ quadrature points. Regarding the number of grid points, two cases are considered i.e., 61 and 101 points corresponding to a maximum error, $|\mu_c - \mu_{true}|/2$, of order 0.1 and 0.06, respectively. To investigate the sensitivity of inference to various shapes of the random effects distribution we have varied the component variance such that $\sqrt{D} = 0.1, 0.4, 0.7, 1$. The fitted random effects densities for these 8 scenarios are shown in Figure 1.

[Figure 1 about here.]

It is evident that increasing dramatically the number of grid points, for the same $\sqrt{D}$ value, does not change considerably the shape of the estimated density. Thus, consideration of a moderate number of grid points, such as 61 for this study, is advisable. Figure 1 shows also that increasing $\sqrt{D}$, for fixed number of points, results in different shapes of the random effects distribution, that vary from an almost discrete (Figure 1a) to a very smooth (Figure 1d) solution. Assuming a discrete random effects distribution is unrealistic, while the oversmoothed solution obscures interesting features in the data. On the contrary an intermediate solution corresponding to Figures 1b and 1c reveals useful features of the data without being neither too rough nor too smooth. Moreover, in Tables 2 and 3 the impact of the various distributional assumptions on the estimated parameters and standard errors is investigated. Different inferences are found for some parameters under different $\sqrt{D}$ values. In particular, for the age effect the p-value, using the Wald test at 5%, varies from 0.03 ($\sqrt{D} = 0.1$) to 0.824 ($\sqrt{D} = 0.4$). Similar conclusions are derived for the covariates BMI, DDays and certain correlation parameters. Regarding the treatment effect, the marginal fitted evolutions are statistically different for at least two groups but the p-value varies from 0.0002 ($\sqrt{D} = 1$) to 0.0130 ($\sqrt{D} = 0.1$).

[Table 2 about here.]

[Table 3 about here.]

As can be seen in Table 2, all models suggest that there is a statistically significant (p-value < 0.001) negative association between the Y and R processes but its size varies from -1.367 (std.err. 0.0812) to -2.156 (std.err. 0.2067). A negative association implies that, the higher the disease activity, the lower is the probability of showing up. The classification of the individuals using the model with 61 grid points and $\sqrt{D} = 0.4$

reveals that the patients are grouped into 3 classes. These classes have a direct interpretation regarding compliance. In particular, one class represents the patients that stayed few days on drug (median 20 days), whereas the rest represent the patients that stayed many days on drug (84 and 87 median days, respectively). Regarding the visiting behavior, the 2 classes represent the completers and the patients that missed at most 2 visits.

Finally, the standard errors presented in Tables 2 and 3 have been estimated, as it has been explained in Section 4.3, using the second derivative matrix of $\ell_n(\theta \mid \hat{G})$ evaluated at $\hat{\theta}$. For the model with $\sqrt{D} = 0.4$, the standard errors have been also estimated using non-parametric Bootstrap with 200 replicates. The standard errors derived by the two methods are similar for many parameter estimates, suggesting that the observed hessian returns quite reliable standard errors.

7. Discussion

A shared parameter model with a flexible random effects distribution has been proposed for analysing longitudinal responses while adjusting for non monotone missingness. The density of the shared latent process has been expressed as a mixture with an unspecified number of normal components, allowing thereby for general shapes of distribution for the random effects. This formulation provides a method for performing sensitivity analysis within the SPM framework by considering various random effects distributions that vary from an almost discrete to an oversmoothed one. For the same dataset tuning the component variance potential skeweness and multimodality is revealed. Thereby more reliable parameter estimates and empirical bayes estimates of better quality than the typical normal distribution can be derived.

Even though our presentation of the proposed model has been concentrated on

the non monotone missingness case, monotone missingness mechanisms can be easily handled by replacing the logistic mixed effects model by an appropriate dropout model. In addition, in our model formulation we have considered only a shared random intercept term. Extensions to random slopes is mathematically straightforward, though more computationally intensive depending on the number of grid and quadrature points used.

The proposed methodology can be easily applied to ignorable mixed effects models as well, allowing thus for flexible random effects distributions. As in the missing data context, the component covariance matrix controls the flexibility of the assumed distribution. However, in the complete data case, the data provide information on the shape of the random effects density and choice of the optimal D is required. Regarding this choice the following suggestions can be made. According to our experience choosing $D \approx \hat{\sigma}_b^Y/4$ in the univariate case, where $\hat{\sigma}_b^Y$ is the estimated variance under the ignorable mixed model with normal random effects, would be in most cases sufficient, whereas larger values may lead to an oversmoothed solution. More formal solutions include the use of the EM algorithm at a second stage or information criteria. In particular, the solution obtained from estimating the SPSP model with $D \approx \hat{\sigma}_b^Y/4$ using the two-step procedure described in Section 4.1 can be used as initial value for the EM algorithm which will then refine D. The use of the EM in this case will not suffer from the convergence difficulties mentioned in Section 3 since a good starting point is provided. Regarding the use of information criteria for choosing the optimal D, further research is required in determining the effective number of parameters which to our knowledge has not yet been investigated in the literature in this context. Finally, the use of the inverse of the second derivative matrix of $\ell_n(\theta \mid \hat{G})$ evaluated at $\hat{\theta}$ has been suggested for the estimation of standard errors. However,

further investigation on the impact of considering the efficient information matrix proposed by van der Vaart (1996) is required.

Acknowledgements

The authors would like to thank Boehringer Ingelheim for permission to use the Rheumatoid Arthritis data. All authors acknowledge support from the Interuniversity Attraction Poles Program P5/24 – Belgian State – Federal Office for Scientific, Technical and Cultural Affairs.

References

Böhning, D. (1985). Numerical estimation of a probability measure. *Journal of Statistical Planning and Inference* **11**, 57–69.

Böhning, D. (1999). *Computer-Assisted Analysis of Mixtures and Applications*. Number 81 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, New York.

Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2004). *Applied Longitudinal Analysis*. Wiley, New Jersey.

Follmann, D. A. and Lambert, D. (1989). Generalizing logistic regression by non parametric mixing. *Journal of the American Statistical Association* **84**, 295–300.

Furst, D. E., Kolba, K. S., Fleischmann, R., Silverfield, J., Greenwald, M., Roth, S., Hall, D. B. and Roszko, P. J. (2002). Dose response and safety study of meloxicam up to 22.5 mg daily in ra: A 12 week multicenter, double blind, dose response study versus placebo and diclofenac. *Journal of Rheumatology* **29**, 436–446.

Jewell, N. P. (1982). Mixtures of exponential distributions. *Annals of Statistics* **10**, 479–484.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics* **27**, 886–906.

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

Lin, H., McCulloch, C. E. and Rosenheck, R. A. (2004). Latent pattern mixture models for informative intermittent missing data in longitudinal studies. *Biometrics* **60**, 295–305.

Lin, H., Turnbull, B. W., McCulloch, C. E. and Slate, E. H. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association* **97**, 53–65.

Lindsay, B. G. (1983). The geometry of mixture likelihoods: A general theory. *The Annals of Statistics* **11**, 86–94.

Magder, L. S. and Zeger, S. L. (1996). A smooth nonparametric estimate of a mixing distribution using mixtures of gaussians. *Journal of the American Statistical Association* **91**, 1141–1151.

Roy, J. (2003). Modeling longitudinal data with nonignorable dropouts using a latent dropout class model. *Biometrics* **59**, 829–836.

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

Simar, L. (1976). Maximum likelihood estimation of a compound poisson process. *Annals of Statistics* **4**, 1200–1209.

Song, X., Davidian, M. and Tsiatis, A. A. (2002). A semiparametric likelihood approach to joint modelling of longitidinal and time to event data. *Biometrics* **58**, 742–753.

van der Vaart, A. W. (1996). Efficient estimation in semiparametric models. *Annals of Statistics* **24**, 862–878.

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

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53**, 330–339.

Appendix A

Derivatives for the SPSP model

Here we present the partial derivatives of the log-likelihood (5) with respect to $\theta = (\beta^{(1)}, \beta^{(2)}, \sigma^2, \alpha^{(1)}, \alpha^{(2)}, \gamma, \mathcal{P}(\rho))$. In particular, we get

$$\begin{aligned}
\frac{\partial \ell}{\partial \beta^{(1)}} &= \sum_{i=1}^{n} \frac{1}{\sigma^2 f(Y_i, R_i; G)} \sum_{h=1}^{H} \varpi_h f(Y_i, R_i \mid b_{ih}; \theta) f(b_{ih}; G) \mathcal{P}_i(\rho)^{-1} X_i^T \eta_{ih}, \\
\frac{\partial \ell}{\partial \beta^{(2)}} &= \sum_{i=1}^{n} \frac{1}{\sigma^2 f(Y_i, R_i; G)} \sum_{h=1}^{H} \varpi_h f(Y_i, R_i \mid b_{ih}; \theta) f(b_{ih}; G) \mathcal{P}_i(\rho)^{-1} Z_i^T \eta_{ih},
\end{aligned}$$

where $\sum_{h=1}^{H}$ corresponds to the Gauss Hermite quadrature rule with H abscissas and corresponding weights ϖ_h, $f(b_{ih}; G) = \sum_{c=1}^{C} \pi_c f(b_{ih}; \mu_c, D)$ is the mixture random effects density and $f(Y_i, R_i \mid b_{ih}; \theta)$ denotes the joint conditional density of the Y and R processes given as

$$f(Y_i, R_i \mid b_{ih}; \theta) = (2\pi)^{-n_i/2} |\Sigma_i|^{-1/2} \exp\{-\frac{1}{2} \eta_{ih}^T \Sigma_i^{-1} \eta_{ih}\} \times \prod_{j=1}^{\mathcal{N}-1} p_{ij}^{r_{ij}} (1 - p_{ij})^{(1-r_{ij})},$$

where $\eta_{ih} = Y_i - X_i \beta^{(1)} - Z_i\ (\beta^{(2)} + b_{ih})$, $\Sigma_i = \sigma^2 \mathcal{P}_i(\rho)$ and $p_{ij} = \text{expit}\{w_{ij}^T \alpha^{(1)} + z_{ij}^T (\alpha^{(2)} + \gamma b_{ih})\}$,

$$\begin{aligned}
\frac{\partial \ell}{\partial \sigma^2} &= \sum_{i=1}^{n} \frac{1}{2\sigma^2 f(Y_i, R_i; G)} \sum_{h=1}^{H} \varpi_h f(Y_i, R_i \mid b_{ih}; \theta) f(b_{ih}; G) [\frac{1}{\sigma^2} \eta_{ih}^T \mathcal{P}_i(\rho)^{-1} \eta_{ih} - p_i], \\
\frac{\partial \ell}{\partial \alpha^{(1)}} &= \sum_{i=1}^{n} \frac{1}{f(Y_i, R_i; G)} \sum_{h=1}^{H} \varpi_h f(Y_i, R_i \mid b_{ih}; \theta) f(b_{ih}; G) \sum_{j=1}^{\mathcal{N}-1} (r_{ij} - p_{ij}) w_{ij}, \\
\frac{\partial \ell}{\partial \alpha^{(2)}} &= \sum_{i=1}^{n} \frac{1}{f(Y_i, R_i; G)} \sum_{h=1}^{H} \varpi_h f(Y_i, R_i \mid b_{ih}; \theta) f(b_{ih}; G) \sum_{j=1}^{\mathcal{N}-1} (r_{ij} - p_{ij}) z_{ij}, \\
\frac{\partial \ell}{\partial \gamma} &= \sum_{i=1}^{n} \frac{1}{f(Y_i, R_i; G)} \sum_{h=1}^{H} \varpi_h f(Y_i, R_i \mid b_{ih}; \theta) f(b_{ih}; G) \sum_{j=1}^{\mathcal{N}-1} (r_{ij} - p_{ij}) z_{ij} b_{ih}, \\
\frac{\partial \ell}{\partial \rho} &= \sum_{i=1}^{n} \frac{1}{2 f(Y_i, R_i; G)} \sum_{h=1}^{H} \varpi_h f(Y_i, R_i \mid b_{ih}; \theta) f(b_{ih}; G) \mathcal{A}_i,
\end{aligned}$$

where $\mathcal{A}_i = -tr\{\mathcal{P}_i^{-1}(\rho) \frac{\partial \mathcal{P}_i(\rho)}{\partial \rho}\} + \frac{1}{\sigma^2} \eta_{ih}^T \mathcal{P}_i^{-1}(\rho) \frac{\partial \mathcal{P}_i(\rho)}{\partial \rho} \mathcal{P}_i^{-1}(\rho) \eta_{ih}$.

Figure 1. Rheumatoid Arthritis study - Estimated random effects density obtained by fitting the SPSP model with varying number of grid points and flexibility. (a) $\sqrt{D} = 0.1$, (b) $\sqrt{D} = 0.4$, (c) $\sqrt{D} = 0.7$, (d) $\sqrt{D} = 1$. The solid and dashed lines correspond to the fit with 61 and 101 equidistant grid points respectively.

Table 1

Simulation study results: Evaluation of the SPSP model and comparison versus the SPM, $n = 50$ *and* $\mathcal{N} = 5$

True Distr	Parms	True Value	SPSP Est (sd) (RMSE)	SP Est (sd) (RMSE)
Normal	β_0	-1.20	-1.17 (0.2943) (0.2952)	-1.19 (0.3302) (0.3300)
	β_1	0.50	0.50 (0.0516) (0.0517)	0.50 (0.0494) (0.0494)
	β_2	-1.50	-1.50 (0.0644) (0.0644)	-1.50 (0.0594) (0.0593)
	α_0	1.60	1.68 (0.6929) (0.6962)	1.67 (0.6990) (0.7013)
	α_1	2.50	2.61 (0.5024) (0.5130)	2.60 (0.4971) (0.5072)
	α_2	-0.50	-0.52 (0.1755) (0.1763)	-0.52 (0.1754) (0.1762)
	γ	0.70	0.73 (0.1438) (0.1472)	0.73 (0.1439) (0.1475)
	σ	0.50	0.49 (0.0307) (0.0313)	0.49 (0.0309) (0.0322)
	σ_b	2.00	1.97 (0.1851) (0.1880)	1.97 (0.1957) (0.1977)
	$\sqrt{\text{ISE}}$		0.12 (0.0314)	0.03 (0.0207)
Normal Mixture	β_0	-1.20	-1.19 (0.3083) (0.3081)	-1.19 (0.3516) (0.3514)
	β_1	0.50	0.49 (0.0481) (0.0484)	0.50 (0.0504) (0.0503)
	β_2	-1.50	-1.50 (0.0522) (0.0522)	-1.50 (0.0623) (0.0623)
	α_0	1.60	1.63 (0.7675) (0.7674)	1.62 (0.7708) (0.7704)
	α_1	2.50	2.58 (0.5093) (0.5156)	2.59 (0.5169) (0.5244)
	α_2	-0.50	-0.51 (0.1926) (0.1929)	-0.51 (0.1911) (0.1912)
	γ	0.70	0.73 (0.1277) (0.1299)	0.73 (0.1304) (0.1342)
	σ	0.50	0.49 (0.0287) (0.0302)	0.49 (0.0303) (0.0319)
	σ_b	2.06	2.10 (0.0915) (0.0986)	2.04 (0.0923) (0.0941)
	$\sqrt{\text{ISE}}$		0.21 (0.0669)	0.42 (0.0065)
t Mixture	β_0	-1.20	-1.20 (0.3163) (0.3160)	-1.17 (0.4111) (0.4121)
	β_1	0.50	0.50 (0.0472) (0.0474)	0.50 (0.0467) (0.0466)
	β_2	-1.50	-1.50 (0.0569) (0.0570)	-1.50 (0.0614) (0.0613)
	α_0	1.60	1.68 (0.7924) (0.7958)	1.70 (0.8171) (0.8219)
	α_1	2.50	2.64 (0.5016) (0.5196)	2.64 (0.5125) (0.5313)
	α_2	-0.50	-0.53 (0.2026) (0.2040)	-0.52 (0.2013) (0.2025)
	γ	0.70	0.72 (0.1288) (0.1309)	0.73 (0.1307) (0.1336)
	σ	0.50	0.50 (0.0369) (0.0372)	0.49 (0.0327) (0.0338)
	σ_b	2.06	2.16 (0.1209) (0.1547)	2.16 (0.1779) (0.2013)
	$\sqrt{\text{ISE}}$		0.18 (0.0640)	0.36 (0.0082)
Log-Normal	β_0	-1.20	-1.38 (0.2299) (0.2901)	-0.75 (1.3311) (1.4026)
	β_1	0.50	0.54 (0.1111) (0.1186)	0.50 (0.0613) (0.0612)
	β_2	-1.50	-1.52 (0.1046) (0.1061)	-1.51 (0.0760) (0.0761)
	α_0	1.60	1.61 (0.7124) (0.7117)	1.96 (1.2031) (1.2553)
	α_1	2.50	2.54 (0.4427) (0.4440)	2.55 (0.4369) (0.4389)
	α_2	-0.50	-0.51 (0.1846) (0.1849)	-0.51 (0.1849) (0.1852)
	γ	0.70	0.75 (0.2354) (0.2354)	0.71 (0.2120) (0.2121)
	σ	0.50	0.55 (0.1542) (0.1617)	0.50 (0.0337) (0.0339)
	σ_b	2.00	1.49 (0.2607) (0.5777)	1.89 (1.0230) (1.0274)
	$\sqrt{\text{ISE}}$		0.28 (0.0455)	0.41 (0.0385)

Table 2

Rheumatoid Arthritis study - Results for the Y and R process: Parameter estimates with standard errors in the brackets calculated using the root of the inverse hessian matrix at convergence. For the $\sqrt{D} = 0.4$ scenario non-parametric bootstrap standard errors are also given in the second bracket.

		Y - Process		
Parms	$\sqrt{D} = 0.1$	$\sqrt{D} = 0.4$	$\sqrt{D} = 0.7$	$\sqrt{D} = 1$
Int	6.877 (0.4250)	5.750 (0.4035) (0.3216)	5.353 (0.3998)	5.496 (0.4035)
t_{ij}	-0.653 (0.0546)	-0.644 (0.0536) (0.0521)	-0.642 (0.0553)	-0.650 (0.0543)
t_{ij}^2	0.043 (0.0042)	0.042 (0.0042) (0.0038)	0.042 (0.0043)	0.043 (0.0042)
Gender_i	-0.226 (0.1310)	-0.137 (0.1270) (0.1241)	-0.139 (0.1260)	-0.070 (0.1261)
Age_i	-0.010 (0.0046)	-0.001 (0.0045) (0.0045)	-0.001 (0.0044)	0.002 (0.0044)
BMI_i	0.013 (0.0082)	0.028 (0.0081) (0.0079)	0.028 (0.0080)	0.031 (0.0080)
DaysD_i	0.002 (0.0019)	0.004 (0.0017) (0.0070)	0.010 (0.0017)	0.002 (0.0017)
$\mathcal{T}_{2i} \times t_{ij}$	-0.007 (0.0715)	-0.027 (0.0720) (0.0756)	-0.020 (0.0744)	-0.034 (0.0724)
$\mathcal{T}_{3i} \times t_{ij}$	0.059 (0.0701)	0.042 (0.0704) (0.0707)	0.038 (0.0727)	0.029 (0.0710)
$\mathcal{T}_{4i} \times t_{ij}$	0.224 (0.0774)	0.242 (0.0763) (0.0731)	0.263 (0.0789)	0.266 (0.0775)
$\mathcal{T}_{5i} \times t_{ij}$	0.160 (0.0734)	0.154 (0.0727) (0.0789)	0.168 (0.0752)	0.171 (0.0737)
$\mathcal{T}_{2i} \times t_{ij}^2$	0.001 (0.0056)	0.002 (0.0056) (0.0056)	0.002 (0.0058)	0.002 (0.0057)
$\mathcal{T}_{3i} \times t_{ij}^2$	-0.006 (0.0055)	-0.005 (0.0055) (0.0053)	-0.005 (0.0057)	-0.004 (0.0056)
$\mathcal{T}_{4i} \times t_{ij}^2$	-0.015 (0.0061)	-0.016 (0.0060) (0.0053)	-0.017 (0.0062)	-0.018 (0.0061)
$\mathcal{T}_{5i} \times t_{ij}^2$	-0.013 (0.0057)	-0.012 (0.0057) (0.0060)	-0.013 (0.0059)	-0.014 (0.0058)
σ^2	4.456 (0.1468)	4.257 (0.1373) (0.5048)	3.843 (0.1639)	4.197 (0.1402)
σ_b^2	1.365	1.383 (0.8451)	2.226	1.415
		R - Process		
Parms	$\sqrt{D} = 0.1$	$\sqrt{D} = 0.4$	$\sqrt{D} = 0.7$	$\sqrt{D} = 1$
Int	4.455 (0.3424)	4.617 (0.4229) (0.5538)	4.457 (0.4365)	4.909 (0.5437)
$\mathcal{T}_{2i}$	-0.744 (0.1688)	-0.826 (0.1893) (0.2180)	-0.831 (0.1888)	-0.903 (0.2008)
$\mathcal{T}_{3i}$	0.038 (0.0184)	0.047 (0.0198) (0.0199)	0.050 (0.0196)	0.055 (0.0208)
$\mathcal{T}_{4i}$	-0.207 (0.4730)	-0.300 (0.5442) (0.5720)	-0.355 (0.5453)	-0.346 (0.6186)
$\mathcal{T}_{5i}$	-0.237 (0.4481)	-0.170 (0.5552) (0.5378)	-0.133 (0.5570)	-0.054 (0.6288)
t_{ij}	0.208 (0.4228)	0.122 (0.5270) (0.5114)	0.142 (0.5271)	0.077 (0.5975)
t_{ij}^2	0.373 (0.4577)	0.405 (0.5821) (0.5500)	0.386 (0.5829)	0.470 (0.6568)
$\mathcal{T}_{2i} \times t_{ij}$	0.021 (0.2394)	0.103 (0.2719) (0.3113)	0.118 (0.2698)	0.152 (0.2864)
$\mathcal{T}_{3i} \times t_{ij}$	-0.138 (0.2333)	-0.062 (0.2783) (0.2770)	-0.041 (0.2764)	-0.019 (0.2927)
$\mathcal{T}_{4i} \times t_{ij}$	-0.714 (0.2203)	-0.717 (0.2717) (0.2866)	-0.667 (0.2665)	-0.745 (0.2837)
$\mathcal{T}_{5i} \times t_{ij}$	-0.155 (0.2411)	-0.176 (0.2887) (0.2924)	-0.176 (0.2857)	-0.199 (0.3029)
$\mathcal{T}_{2i} \times t_{ij}^2$	-0.007 (0.0259)	-0.012 (0.0285) (0.0297)	-0.013 (0.0282)	-0.016 (0.0299)
$\mathcal{T}_{3i} \times t_{ij}^2$	0.021 (0.0254)	0.014 (0.0292) (0.0265)	0.012 (0.0288)	0.010 (0.0306)
$\mathcal{T}_{4i} \times t_{ij}^2$	0.077 (0.0243)	0.076 (0.0284) (0.0275)	0.072 (0.0277)	0.079 (0.0295)
$\mathcal{T}_{5i} \times t_{ij}^2$	0.005 (0.0263)	0.006 (0.0299) (0.0286)	0.008 (0.0294)	0.008 (0.0312)
γ	-1.767 (0.0812)	-1.820 (0.1298) (0.4526)	-1.367 (0.1176)	-2.156 (0.2067)
log-Lik	-8422.377	-8929.576	-8966.621	-8947.095

Table 3

Rheumatoid Arthritis study - Results for the correlation matrix: Parameter estimates with standard errors in the brackets calculated using the root of the inverse hessian matrix at convergence. For the $\sqrt{D} = 0.4$ scenario non-parametric bootstrap standard errors are also given in the second bracket.

	Correlation			
Parms	$\sqrt{D} = 0.1$	$\sqrt{D} = 0.4$	$\sqrt{D} = 0.7$	$\sqrt{D} = 1$
ρ_{21}	0.239 (0.0487)	0.212 (0.0455) (0.1315)	0.061 (0.0599)	0.178 (0.0486)
ρ_{31}	0.195 (0.0472)	0.169 (0.0461) (0.1351)	0.019 (0.0609)	0.135 (0.0489)
ρ_{32}	0.589 (0.0259)	0.579 (0.0258) (0.0748)	0.508 (0.0342)	0.559 (0.0278)
ρ_{41}	0.249 (0.0445)	0.221 (0.0441) (0.1279)	0.086 (0.0577)	0.188 (0.0469)
ρ_{42}	0.358 (0.0389)	0.331 (0.0399) (0.1178)	0.216 (0.0527)	0.302 (0.0427)
ρ_{43}	0.567 (0.0274)	0.553 (0.0280) (0.0853)	0.480 (0.0368)	0.532 (0.0302)
ρ_{51}	0.107 (0.0386)	0.077 (0.0389) (0.1361)	-0.056 (0.0489)	0.042 (0.0405)
ρ_{52}	0.364 (0.0327)	0.343 (0.0329) (0.1106)	0.242 (0.0425)	0.317 (0.0346)
ρ_{53}	0.471 (0.0277)	0.455 (0.0282) (0.0912)	0.377 (0.0363)	0.433 (0.0298)
ρ_{54}	0.595 (0.0256)	0.583 (0.0262) (0.0732)	0.527 (0.0329)	0.564 (0.0283)