Thesis mid-term report

By: Georgy Gomon
Studentnumber: s1559370
Date: November 18, 2021

Abstract

In longitudinal data analysis one often encounters endogenous time-dependent covariates: these are covariates whose current value, given their own history, depends on past values of the outcome. In the presence of such endogenous covariates, because of the cross-reliance of the endogenous covariate on the outcome, standard Mixed Models are no longer valid and one needs to resort to joint modelling of both the outcome and the endogenous covariate. In this thesis several such joint longitudinal models will be discussed. To fit these models we shall be examining a novel Bayesian technique called INLA (Integrated Nested Laplace Integration), which is a simple technique that could possibly replace the complex and long MCMC estimation procedure. Although INLA has seen rapid development over the past years, joint longitudinal models have so far received little attention. The goal of this thesis is to implement several joint longitudinal models within the INLA framework and apply them to a LUMC dataset that will be chosen at a later date.

Contents

1 Introduction

Longitudinal data analysis focuses on the effect covariates have on a certain outcome over time. As an example we can imagine studying the effect of the covariates 'sex', 'age' and 'treatment regime' on the outcome 'lung capacity' following a COVID infection. Within the longitudinal framework we would then measure the values of the covariates and outcome multiple times over the span of e.g. a few years. Within the context of longitudinal data we can split the covariates into 3 groups. Covariates can be time-dependent or time-independent. In our hypothetical example 'sex' is a time-independent covariate, as it does not change over time. The time dependent covariates can be split into endogenous and exogenous time-dependent covariates. An exogenous time-dependent covariate is a covariate whose current value, given its own history, does not depend on the value of the outcome at previous measurement times. In our example 'age' is such an exogenous time-dependent covariate. 'Age' does change over time, but it is independent of 'lung capacity' (the outcome) at previous time points. Lastly, we have the endogenous time-dependent covariates, which are covariates whose current value does depend on previous values of the outcome, given their own history. In our example 'treatment regimen' is such an endogenous time-dependent covariate, since the lung capacity at previous measurements can influence the treatment regimen the patient is currently receiving, e.g: If the patient is recovering the treatment can be scaled down. Modelling such endogenous time-dependent covariates (we shall call them endogenous covariates) is difficult, since there is a causal path from outcome to endogenous covariate and vice versa. A standard linear mixed model is no longer applicable but instead the endogenous covariate and the outcome need to be modelled jointly. This leads us into the framework of joint longitudinal models. For more information on endogenous covariates and joint models we refer to [P.J. Diggle, 2016]. Within the scope of this thesis the different approaches to joint modelling of the endogenous covariate and the outcome will be studied. The emphasis will be on 3 methods:

- First is a simple multivariate model in which the multiple outcomes are jointly Gaussian distributed. The association between the two outcomes is then modelled via correlated errors in the linear predictors of the outcomes.
- Second is a joint mixed model in which the association between the multiple outcomes is given by multivariate normally distributed random effects and multivariate normally distributed errors terms.
- Lastly a joint model is proposed in which the linear predictor of the endogenous covariate is inserted into the linear predictor of the outcome with an associated scaling factor.

During the Thesis these methods will be applied in R within the Bayesian framework. The emphasis will be to implement the methods using INLA (Integrated Nested Laplace Approximation) and its associated R package R-INLA. INLA is a new Bayesian framework based on Laplace Integration that removes the need for extensive MCMC estimation and is therefore much quicker than standard Bayesian methods. For more information on INLA we refer to [Rue et al., 2009]. For more information about the current joint models implementations of INLA we refer to [van Niekerk et al., 2021].

2 Joint Mixed Models

Within longitudinal studies we have both time-invariant and time-varying covariates. Examples of time-invariant covariates are sex, treatment group and genetic profile. Examples of time-varying covariates are age, biomarkers, air-pollution exposure and treatment dose. The time-varying covariates can furthermore be divided into 2 groups: Exogenous and Endogenous covariates.
Before introducing the difference between these covariates some mathematical notation will be introduced:

- $y_i(t)$: Value of the response y for subject i at time t.
- $x_i(t)$: Value of the covariate x for subject i at time t.
- $\mathcal{H}_i^Y$: History of the response process of subject i until time t:

$$\mathcal{H}_i^Y(t) = \{y_i(t_{i1}), y_i(t_{i2}), ..., y_i(t_{ik}); t_{ik} < t\}$$

- $\mathcal{H}_i^X$: History of the covariate process of subject i until time t:

$$\mathcal{H}_i^X(t) = \{x_i(t_{i1}), x_i(t_{i2}), ..., x_i(t_{ik}); t_{ik} < t\}$$

- $\mathbf{W}_i$: Vector of time-independent covariates.

2.1 Endogenous vs Exogenous covariates

Definition 2.1 (Exogenous Covariate) *$X_i(t)$ is an exogenous covariate with respect to the outcome process if the exposure at time t is conditionally independent on the history of the outcome process at time t, given the history of the exposure process at time t. Mathematically,*

$$f\left(x_i(t)|\mathcal{H}_i^Y(t), \mathcal{H}_i^X(t-1), \mathbf{W}_i\right) = f\left(x_i(t)|\mathcal{H}_i^X(t-1), \mathbf{W}_i\right)$$

Thus, for an exogenous covariate the exposure at time t does not depend on previous values of the response. Examples of exogenous covariates are age and air-pollution exposure.
For exogenous covariates the likelihood $f\left(\mathbf{Y_i}, \mathbf{X_i}|\mathbf{W_i}, \theta\right)$ can be factorized:

$$\begin{aligned} f\left(\mathbf{Y_i}, \mathbf{X_i}|\mathbf{W_i}, \theta\right) &= \left[\prod_{t=1}^{T} f\left(y_i(t)|\mathcal{H}_i^Y(t-1), \mathcal{H}_i^X(t), \mathbf{W}_i, \theta\right)\right] \cdot \left[\prod_{t=1}^{T} f\left(x_i(t)|\mathcal{H}_i^X(t-1), \mathbf{W}_i, \theta\right)\right] = \\ &= \mathcal{L}_Y(\theta_1) \cdot \mathcal{L}_X(\theta_2) \end{aligned}$$

The factorization of the joint likelihood means that we do not need to model the covariate process of X in order to make inference about θ_1 and the outcome Y.

Definition 2.2 (Endogenous Covariate) *$X_i(t)$ is an endogenous covariate with respect to the outcome process if the exposure at time t is conditionally dependent on the history of the outcome process at time t, given the history of the exposure process at time t. Mathematically,*

$$f\left(x_i(t)|\mathcal{H}_i^Y(t), \mathcal{H}_i^X(t-1), \mathbf{W}_i\right) \neq f\left(x_i(t)|\mathcal{H}_i^X(t-1), \mathbf{W}_i\right)$$

Thus, for an exogenous covariate the exposure at time t does depend on previous values of the response. An example might occur in case of a non-controlled study of the effect of a certain treatment regimen on symptom severity. If no symptoms are present, the treatment regimen might be made less stringent and vice-versa.
For exogenous covariates the factorization shown above can not be done, and thus the joint process of X and Y needs to be modelled in order to make inference about Y.

2.2 Mixed Models

We shall be using Mixed Models to analyse the longitudinal data. The mixed models are of the following form:

$$y_i(t_{ij}) = \mathbf{w}_i \cdot \alpha + \mathbf{x}_i(t_{ij}) \cdot \beta + \mathbf{z}_i(t_{ij}) \cdot \mathbf{b_i} + \epsilon_i(t_{ij})$$

with:

- $y_i(t_{ij})$: Outcome for patient i at time t_{ij}.
- $\mathbf{w}_i$: Vector of fixed time-independent covariates
- $\mathbf{x}_i(t_{ij})$: Vector of fixed time-varying covariates at time t_{ij}.
- α and β: Coefficients of fixed time-independent and time-varying covariates respectively
- $\mathbf{z}_i(t_{ij})$: Vector of Random time-independent covariates
- $\mathbf{b_i} \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$: Random effects vector
- $\epsilon_i(t_{ij}) \sim \mathcal{N}(0, \sigma^2)$: Residual errors, with $\epsilon_i(t_{ij}) \perp\!\!\!\perp \mathbf{b_i}$ and $\mathbf{b_i} \perp\!\!\!\perp \mathbf{x}_i(t_{ij}), \mathbf{w}_i$.

2.3 Joint Mixed Models

As was shown in definition 2.2, in case of endogenous covariates the likelihood $f\left(\mathbf{Y_i}, \mathbf{X_i}|\mathbf{W_i}, \theta\right)$ of the outcome Y and endogenous covariate X can not be factorized and thus both the endogenous covariate X and the outcome Y need to be modelled jointly to make inference.
In this thesis we shall be looking at 3 main models which jointly model the outcome and the endogenous covariate.

2.3.1 Multivariate Joint Model

The first joint mixed model we shall be examining is a multivariate normal joint mixed model. Here the association between the outcome Y and the endogenous covariate X is supplied via the residual errors covariance matrix $\mathbf{\Sigma}_i$.

$$\begin{cases} y_i(t_{ij}) = \mathbf{v}_{yi}^\intercal(t_{ij})\beta_{\mathbf{y}} + \epsilon_{yi}(t_{ij}) \\ x_i(t_{ij}) = \mathbf{v}_{xi}^\intercal(t_{ij})\beta_{\mathbf{x}} + \epsilon_{xi}(t_{ij}) \end{cases} \quad \text{with} \quad \begin{bmatrix} \epsilon_{yi} \\ \epsilon_{xi} \end{bmatrix} \sim \mathcal{N}_{n_i}(\mathbf{0}, \mathbf{\Sigma}_i)$$

In this model the association can be measured between any pair of time-points and missing data in the response or covariates can be handled simultaneously.
The largest disadvantage of this method is that it allows only for balanced designs and that all covariates and outcomes must be measured at the same time point.
A choice must be made for the structure of the variance-covariance matrix $\mathbf{\Sigma}_i$. Possible choices are an unstructured form (however requiring many parameters to be estimated), compound symmetry, Auto-regressive and Toeplitz.

2.3.2 Joint Mixed Model

The next type of models we shall be examining are joint mixed models of the form shown below:

$$\begin{cases} y_i(t_{ij}) = \mathbf{v}_{yi}^\intercal(t_{ij})\beta_{\mathbf{y}} + \mathbf{z}_{yi}^\intercal(t_{ij})\mathbf{b}_{yi} + \epsilon_{yi}(t_{ij}) \\ x_i(t_{ij}) = \mathbf{v}_{xi}^\intercal(t_{ij})\beta_{\mathbf{x}} + \mathbf{z}_{xi}^\intercal(t_{ij})\mathbf{b}_{xi} + \epsilon_{xi}(t_{ij}) \end{cases} \quad \text{with}$$

$$\begin{bmatrix} \mathbf{b}_{yi} \\ \mathbf{b}_{xi} \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{D}); \qquad \begin{bmatrix} \epsilon_{yi} \\ \epsilon_{xi} \end{bmatrix} \sim \mathcal{N}_{n_i}(\mathbf{0}, \mathbf{\Sigma}_i); \qquad \epsilon_{yi}(t_{ij}) \perp\!\!\!\perp \mathbf{b}_{yi}, \epsilon_{xi}(t_{ij}) \perp\!\!\!\perp \mathbf{b}_{xi}$$

Here association is measured via the random effects ($\mathbf{D}$) and the residual errors ($\mathbf{\Sigma}_i$). A large advantage of this model over the Multivariate Joint Model is that here the outcome and endogenous covariate do not need to be measured at the same time, if one does not incorporate association via the residuals errors $\mathbf{\Sigma}_i$.

2.3.3 Mixed Model with scaled linear predictor

Lastly we have a mixed model in which the linear predictor of the endogenous covariate is copied into the linear predictor of the outcome with an associated scaling factor γ.
The model is of the following form:

$$\begin{cases} x_i(t_{ij}) = m_i(t_{ij}) + \epsilon_{xi}(t_{ij}) \\ y_i(t_{ij}) = \mathbf{w}_{yi}^\intercal\alpha_{\mathbf{y}} + \gamma \cdot m_i(t_{ij}) + \mathbf{v}_{yi}^\intercal(t_{ij})\beta_{\mathbf{y}} + \mathbf{z}_{yi}^\intercal(t_{ij})\mathbf{b}_{yi} + \epsilon_{yi}(t_{ij}) \end{cases}$$

with

$$m_i(t_{ij}) = \mathbf{w}_{xi}^\intercal\alpha_{\mathbf{x}} + \mathbf{v}_{xi}^\intercal(t_{ij})\beta_{\mathbf{x}} + \mathbf{z}_{xi}^\intercal(t_{ij})\mathbf{b}_{xi}$$

and

$$\begin{aligned} &\mathbf{b}_{xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{D}_x), \qquad \mathbf{b}_{yi} \sim \mathcal{N}(\mathbf{0}, \mathbf{D}_y) \\ &\epsilon_{yi}(t_{ij}) \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma_y^2), \qquad \epsilon_{xi}(t_{ij}) \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma_x^2) \\ &\epsilon_{yi}(t_{ij}) \perp\!\!\!\perp \mathbf{b}_{yi}, \qquad \epsilon_{xi}(t_{ij}) \perp\!\!\!\perp \mathbf{b}_{xi} \end{aligned}$$

In the above example the linear predictor of X, $m_i(t_{ij})$, is copied with scaling factor γ only at time-point t_{ij}. However, this dependence can be elaborated and linear predictors at any time point $t_{ik} < t_{ij}$ can be included into the linear predictor of the outcome, in this way creating a lagged functional form.

3 INLA

In this thesis we shall be fitting the joint models using Integrated Nested Laplace Approximation (INLA) and it's implementation in R with the R-package R-INLA [Rue et al., 2009]. First a quick overview of INLA will be given, and it will be shown how proposed joint models can be rewritten into the INLA framework. INLA combines the usage of Latent Gaussian Models (LGM's), Gaussian Markov Random Fields (GMRF's), Numerical methods for sparse matrices and Laplace approximations to derive approximate Bayesian inference. Overall INLA is much faster than Monte Carlo Markov Chain (MCMC) methods for Bayesian inference without the need for long sampling chains. Also, research has shown that in terms of accuracy INLA is not inferior to MCMC methods.

3.1 Bayesian Inference using INLA

3.1.1 Latent Gaussian Model

The basis of INLA relies on the fact that many statistical models, including the joint longitudinal models discussed in this thesis, can be rewritten as a Latent Gaussian Model (LGM). Furthermore, only models that can be rewritten as LGM's can be used within the INLA framework.
An LGM consists of the following elements:

- Likelihood of the outcome: $\mathbf{y}|\mathbf{x}, \theta_2 \sim \prod_i p(y_i|\eta_i, \theta_2)$
- The Latent Field: $\mathbf{x}|\theta_1 \sim p(\mathbf{x}, \theta_1) = \mathcal{N}(0, \mathbf{\Sigma})$.
- The Hyperpriors: $\theta = (\theta_1, \theta_2) \sim p(\theta)$.

Here $\mathbf{y}$ is the observed data and $\mathbf{x}$ are all the parameters in the linear predictor. Note that the dimension of $\mathbf{x}$ is usually very large (model has many data-points), but the dimension of θ is usually small (just a few parameters are needed to define the random effects structure).
Imagine we have the most general form of a generalized linear mixed model:

$$y \sim \prod_i^N p(y_i|\mu_i) \quad \text{with} \quad g(\mu_i) \equiv \eta_i = \alpha + \sum_{k=1}^{n_\beta} \beta_k \cdot z_{k_i} + \sum_{j=1}^{n_f} f^{(j)}(w_{ji}) + \epsilon_i$$

Here $g()$ is the link function, α the intercept, β the regression parameters of covariates z and $f()$ the random effects of covariates w.
Such a model is an LGM if and only if we assume that all parameters have a joint Normal distribution, thus:

$$\mathbf{x} = [\eta, \alpha, \beta, f()] \sim \mathcal{N}(0, \Sigma)$$

If we furthermore assume conditional independence in $\mathbf{x}$, then this latent field $\mathbf{x}$ is a Gaussian Markov Random Field.

3.1.2 Gaussian Markov Random Field (GMRF)

Our vector $\mathbf{x} = [\eta, \alpha, \beta, f()] \sim \mathcal{N}(0, \Sigma)$ can then be thought of as a Gaussian Markov Random Field (GMRF). A GMRF is a normally distributed random vector $\mathbf{x} = (x_1, ..., x_n)$ with Markov properties, such as that for some $i \neq j$, $x_i \perp\!\!\!\perp x_j|\mathbf{x}_{-ij}$, which means that x_i is independent of x_j given all elements of $\mathbf{x}$ other than i and j ($\mathbf{x}_{-ij}$). The Markov properties are given in the Precision matrix $Q = \Sigma^{-1}$, which is the inverse of the covariance matrix. Rue et all [Rue et al., 2009] showed that $x_i \perp\!\!\!\perp x_j|\mathbf{x}_{-ij}$ iff $Q_{ij} = 0$. This result ensures that if in our vector $\mathbf{x} = [\eta, \alpha, \beta, f()] \sim \mathcal{N}(0, \Sigma)$ the different elements are independent, the precision matrix Q will be very sparse, allowing for easy and fast computations.

3.1.3 Laplace Approximation

INLA uses the Laplace Approximation to estimate any distribution $g(x)$ with a normal distribution. The first 3 terms of the Taylor expansion around the mode ($\hat{x}$) are used to approximate $\log g(x)$ by:

$$\log g(x) \approx \log g(\hat{x}) + \frac{\delta \log g(\hat{x})}{\delta x}(x - \hat{x}) + \frac{\delta^2 \log g(\hat{x})}{2\delta x^2}(x - \hat{x})^2$$

Now, the second term in the approximation, $\frac{\delta \log g(\hat{x})}{\delta x}(x - \hat{x})$, equals 0, since we are considering the derivative at the mode which is a maximum of the function.
We now estimate the variance as:

$$\hat{\sigma}^2 = -\frac{2\delta x^2}{\delta^2 \log g(\hat{x})}\bigg|_{\hat{x}}$$

Using this we obtain:

$$\log g(x) \approx \log g(\hat{x}) - \frac{1}{2\sigma^2}(x - \hat{x})^2$$

With the last expression we can perform a normal approximation:

$$\int g(x)dx = \int \exp\left[\log g(x)\right] dx \approx \int \exp\left[\log g(\hat{x}) - \frac{1}{2\sigma^2}(x-\hat{x})^2\right] dx =$$
$$= \exp\left[\log g(\hat{x})\right] \cdot \int \exp\left[-\frac{1}{2\sigma^2}(x-\hat{x})^2\right] dx = \text{constant} \cdot \int \exp\left[-\frac{1}{2\sigma^2}(x-\hat{x})^2\right] dx$$

Thus, the distribution of $g(x)$ is now approximated by a normal distribution with mean $\hat{x}$, which is found by solving $g'(x) = 0$ and with variance $\hat{\sigma}^2 = -\frac{2\delta x^2}{\delta^2 \log g(\hat{x})}\Big|_{\hat{x}}$, obtained at the mode $\hat{x}$.

3.1.4 Approximating the Latent Field

When conducing Bayesian inference we are interested in the marginals for the elements of the latent field (e.g: regression coefficients):

$$p(x_i|\mathbf{y}) = \int p(x_i, \theta|\mathbf{y})d\theta = \int p(x_i|\theta, \mathbf{y})p(\theta|\mathbf{y})d\theta$$

and the elements of the hyperprior distribution (e.g: variances of random effects):

$$p(\theta_k|\mathbf{y}) = \int p(\theta|\mathbf{y})d\theta_{-k}$$

To obtain these estimates we need to approximate $p(x_i|\theta, \mathbf{y})$ and $p(\theta|\mathbf{y})$.

3.1.5 Approximating $p(\theta|\mathbf{y})$

We can approximate the marginal distribution as:

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{x}, \theta|\mathbf{y})}{p(\mathbf{x}|\theta, \mathbf{y})} \approx \frac{p(\mathbf{y}|\mathbf{x}, \theta)p(\mathbf{x}|\theta)p(\theta)}{\tilde{p}(\mathbf{x}|\theta, \mathbf{y})}\Bigg|_{x=x^*(\theta)} = \tilde{p}(\theta|\mathbf{y}).$$

Here a Gaussian Laplace approximation is used for the denominator $p(\mathbf{x}|\theta, \mathbf{y})$ at the mode $x = x^*(\theta)$.

3.1.6 Approximating $p(x_i|\theta, \mathbf{y})$

To approximate $p(x_i|\theta, \mathbf{y})$ INLA has 3 options:

- Normal approximation, can be used in INLA when selecting the option 'Gaussian'. Here we approximate $p(x_i|\theta, \mathbf{y})$ using standard Laplace approximation, and because we already computed $\tilde{p}(\mathbf{x}|\theta, \mathbf{y})$ during the exploration of $p(\theta|\mathbf{y})$ only the marginals are left to be computed. This method is by far the fastest of the three but often yields poor results.
- Laplace approximation, used in INLA when selecting the option 'Laplace'. Partitions the latent field $\mathbf{x} = [x_j, \mathbf{x}_{-j}]$ and uses Laplace approximation for each element x_j in the latent field:

  $$p(x_j|\theta, \mathbf{y}) \propto \frac{p(\mathbf{x}, \theta|\mathbf{y})}{p(\mathbf{x}_{-j}|x_j, \theta, \mathbf{y})} \propto \frac{p(\theta)p(\mathbf{x}|\theta)p(\mathbf{y}|\mathbf{x})}{p(\mathbf{x}_{-j}|x_j, \theta, \mathbf{y})}$$

  Overall gives good results because the conditionals $p(\mathbf{x}_{-j}|x_j, \theta, \mathbf{y})$ are often close to normal, but is computationally expensive.
- Simplified Laplace approximation, default setting in INLA. Uses a compromise between the first 2 methods. Is computationally fast and almost always gives results very similar to the Laplace approximation.

For more information regarding Bayesian Inference with INLA we refer to [Rue et al., 2009].

3.2 Model Assessment in INLA

Several methods are implemented in INLA to assess the goodness of fit of a model.

3.2.1 Marginal Likelihood

The Marginal Likelihood, also called Model evidence, is the probability that the data observed comes from a given model, independent of the parameters of that model (the parameters of the model are integrated out). The Marginal Likelihood is a very convenient exclusively Bayesian model assessment tool which enables the comparison between models.
In INLA the Marginal Likelihood is approximated as:

$$\widetilde{\pi}(y) = \int \left. \frac{\pi(\theta, x, y)}{\widetilde{\pi}_G(x|\theta, y)} \right|_{x=x^*(\theta)} d\theta$$

Here $\widetilde{\pi}_G(x|\theta, y)$ is the Gaussian approximation (see section 3.1.3) at the mode $x = x^*(\theta)$.
When considering a set of M models $\{\mathcal{M}_m\}_{m=1}^{M}$, the marginal likelihoods are written down as $\pi(y|\mathcal{M}_m)$. If supplying each model with a prior $\pi(\mathcal{M}_m)$, posterior probabilities for each of the models can be calculated as:$\pi(\mathcal{M}_m|y) \propto \pi(y|\mathcal{M}_m)\pi(\mathcal{M}_m)$. These posteriors can now be used to compute the Bayes factor K for 2 different models $\mathcal{M}_1$ and $\mathcal{M}_2$:

$$K = \frac{\pi(\mathcal{M}_1|y)}{\pi(\mathcal{M}_2|y)} = \frac{\pi(y|\mathcal{M}_1)\pi(\mathcal{M}_1)}{\pi(y|\mathcal{M}_2)\pi(\mathcal{M}_2)}$$

In case of equal priors for the 2 models (the models are considered equally likely), the Bayes factor is simply the fraction of the Marginal Likelihoods of the models: $K = \frac{\pi(\mathcal{M}_1|y)}{\pi(\mathcal{M}_2|y)} = \frac{\pi(y|\mathcal{M}_1)}{\pi(y|\mathcal{M}_2)}$

3.2.2 Conditional Predictive Ordinates (CPO)

The Conditional Predictive Ordinate (CPO) is computed for each observation i as:

$$CPO_i = \pi(y_i|y_{-i}).$$

It is the posterior probability of observing observation y_i when the model is fit using all data but y_i. A small value for an observation might indicate a possible outlier. INLA approximates this quantity for every measurement without the need the re-analyse the model while the observation is removed.
The CPO can be summarized over all the data by:

$$CPO = -\sum_{i=1}^{N} \log(CPO_i)$$

A smaller value indicates a better fit of the model over all observations.

3.2.3 Probability Integral Transform (PIT)

The Probability Integral Transform (PIT) is very similar to the CPO and is computed for each observation as:

$$PIT_i = \pi(y_i^{new} \leq y_i|y_{-i})$$

The PIT measures the probability for a new observation y_i^{new} to be lower than y_i when model is fit using all data but y_i. Both the CPO and PIT are thus a sort of Leave-One-Out Cross-Validation (LOO CV). A very large or small PIT for a given value indicates a possibly surprising observation.
Over all the observations, in case of a good model, the PIT's should be approximately uniformly distributed on $[0, 1]$. The Kolmogorov Smirnov non-parametric test is used to test whether the PIT's are indeed uniformly distributed.

3.2.4 DIC and WAIC

The DIC (Deviance Information Criteria) is a popular method for model selection, as it combines goodness of fit with penalization of the number of parameters used. The DIC is given by:

$$DIC = D(\hat{x}, \hat{\theta}) + 2p_D$$

Here $D(\hat{x}, \hat{\theta})$ is the model deviance, which is calculated using the posterior mean $\hat{x}$ and the posterior mode $\hat{\theta}$, as the distribution of θ can be severely skewed.
The effective number of parameters p_D is approximated as:

$$p_D(\theta) \approx n - tr\{Q(\theta)Q * (\theta)^{-1}\}$$

With n being the number of observations and Q being the precision of the Gaussian Markov Random Field 3.1.2.
The Watanabe-Akaike Information Criterion is similar to the DIC, with the only difference being that the effective number of parameters p_D is calculated in a different way.

3.2.5 Mean Squared Error (MSE)

The last method via which we shall assess goodness of fit and compare models is by using Mean Squared Error (MSE). MSE is not incorporated into the INLA package but was calculated using the posterior means of fitted values. The MSE is given by

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Here N is the total number of measurements while y_i and $\hat{y}_i$ are the actual and fitted (posterior means) outcomes respectively. In order to asses both the marginal and hierarchical model fit 3 types of MSE were calculated:

- MSE on the training set: In order to test how well the model can fit the training subjects and to discover possible instances of overfitting.
- MSE on subsequent measurements of subjects whose random effects have been determined by previous measurements, thus giving hierarchical results. In this way one can test how well the model is able to fit the random effects of each individual.
- MSE on test subjects. Here the interest is only on the marginal results, only showing the ability of the model to correctly estimate fixed effects.

3.3 Priors in INLA

The prior for fixed effects in INLA is a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, in which both the mean and precision $\tau = 1/\sigma^2$ can be specified. The default values supplied by INLA are $\mu = 0, \tau = 0.001$.

3.3.1 Random Effect Priors

Within the scope of this thesis we shall mainly be using independent and identically distributed random effects structures. Imagine we have two random effects, u and v, which are together i.i.d. distributed bivariate Normals:

$$\begin{bmatrix} u \\ v \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{W}^{-1}), \quad \text{with covariance matrix} \quad \mathbf{W}^{-1} = \begin{pmatrix} 1/\tau_u & \rho/\sqrt{\tau_u \tau_v} \\ \rho/\sqrt{\tau_u \tau_v} & 1/\tau_v \end{pmatrix}$$

Here τ_u, τ_v (marginal precisions) and ρ (correlation coefficient) are hyperparameters.
The hyperparameters are represented internally in INLA as $\theta = (\log \tau_u, \log \tau_v, \phi)$, with $\rho = 2\frac{\exp(\phi)}{\exp(\phi)+1} - 1$.
As we are more interested in the variances σ_u^2 and σ_v^2 rather than the precisions τ_u and τ_v we use the inverse of the posterior marginal distribution of the precisions to obtain the corresponding distributions of the variances.
The precision matrix $\mathbf{W}$ is Wishart distributed of $p = 2$ dimensions with support n:

$$\mathbf{W} \sim Wishart_2(n, \mathbf{R}^{-1}) \quad \text{with} \quad \mathbf{R} = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix} \quad \text{and} \quad R_{12} = R_{21} \quad \text{due to symmetry}$$

Some properties of the Wishart distribution are:

$$\mathbb{E}(\mathbf{W}) = n\mathbf{R}^{-1}, \quad \mathbb{E}(\mathbf{W}^{-1}) = \frac{\mathbf{R}}{n - (p+1)}$$

The variance of the Wishart distribution has no easy overall form, but in general the variance is larger with increasing support n.
The iid random effects thus have prior-parameters $(n, R_{11}, R_{21} = R_{12}, R_{22})$. These can be specified with default values (for the case $p = 2$) being $(4, 1, 1, 0)$.

4 Configuring the Joint Models in R-INLA

We started the thesis by configuring the Joint Models in the R-INLA package and testing them on simulated data. During this testing phase we also compared the results obtained using R-INLA with results obtained using the R-packages nlme, lmer and MCMCglmm.

4.1 Independent Mixed Models

The first model we used to test the implementation of Joint Mixed Models in R-INLA was a simple Mixed Model without any association between the 2 outcomes. Throughout the following this shall be seen as a baseline model to compare the other models to, as this model does not capture the association between the endogenous covariate and the outcome, and thus would theoretically not be able to fit well in the presence of a endogenous covariate.

Definition 4.1 (Model 0) *We shall be referring to this model, without any association between the endogenous covariate and the outcome, as Model 0. Here we have 2 independent mixed models, for the endogenous covariate y_1 and the outcome y_2, with in each mixed model a dependent random intercept and random slope. We assume here the Conditional Independence Assumption, the assumption that the random effects capture all correlation between measurements on a patient, and thus no correlation is left for the errors.*
Mathematically, the model is specified as:

$$\begin{cases} y^1_{i,j} = (\beta^1_0 + u^1_{0,i}) + \beta^1_x \cdot x_{i,j} + \left(\beta^1_t + u^1_{t,i}\right) \cdot t_{i,j} + \epsilon^1_{i,j} \\ y^2_{i,j} = (\beta^2_0 + u^2_{0,i}) + \beta^2_x \cdot x_{i,j} + \left(\beta^2_t + u^2_{t,i}\right) \cdot t_{i,j} + \epsilon^2_{i,j} \end{cases} \quad \textit{with}$$
$$\begin{bmatrix} u^1_{0,i} \\ u^1_{t,i} \end{bmatrix} \sim \mathcal{N}_2 \left[\mathbf{0}, \begin{pmatrix} \sigma^2_{1,0} & \sigma_{1,(0,t)} \\ \sigma_{1,(t,0)} & \sigma^2_{1,t} \end{pmatrix} \right]; \quad \begin{bmatrix} u^2_{0,i} \\ u^2_{t,i} \end{bmatrix} \sim \mathcal{N}_2 \left[\mathbf{0}, \begin{pmatrix} \sigma^2_{2,0} & \sigma_{2,(0,t)} \\ \sigma_{2,(t,0)} & \sigma^2_{2,t} \end{pmatrix} \right]$$
$$\begin{bmatrix} \boldsymbol{\epsilon}^1_i \\ \boldsymbol{\epsilon}^2_i \end{bmatrix} \sim \mathcal{N}_{2j}(\mathbf{0}, \mathbf{I}_{2j})$$

Here we use the following notation (which will also be used for the remainder of this chapter):

- $y^1_{i,j}$ *and* $y^2_{i,j}$*: The endogenous covariate and outcome respectively for patient* $i = 1, ..., N$ *at time-point* $j = 1, ..., n$.
- β^1_0, β^1_x & β^1_t*: The fixed effects for the intercept, covariate* $x_{i,j}$ *and time* $t_{i,j}$ *(taken to be linear) for the endogenous covariate* y_1*.* β^2_0, β^2_x & β^2_t *have similar roles for the outcome* y_2*.*
- $u^1_{0,i}$ & $u^1_{t,i}$*: Random intercept and random (time)-slope for patient i for endogenous covariate* y_1*. Together they are normally distributed with mean* 0 *and covariance matrix* $\begin{pmatrix} \sigma^2_{1,0} & \sigma_{1,(0,t)} \\ \sigma_{1,(t,0)} & \sigma^2_{1,t} \end{pmatrix}$.
  $u^2_{0,i}$ & $u^2_{t,i}$*: Are similarly the Random Intercept and Random slope for the outcome* y_2*.*
- $\epsilon^1_{i,j}$ & $\epsilon^2_{i,j}$ *are the errors for the endogenous covariate* y_1 *and the outcome* y_2 *respectively.*

To fit this multiple-likelihood model within the R-INLA framework a few tricks have to be used:

- The response variables $\mathbf{y}^1$ and $\mathbf{y}^2$, as well as the fixed effects (Intercept, covariate $\mathbf{x}$ as well as time $\mathbf{t}$), have to be stored in matrices of the following form:

$$\begin{array}{cccc}
\text{response variables } \mathbf{y}^1 \text{ and } \mathbf{y}^2 & \text{Intercept} & \text{Covariate } \mathbf{x} & \text{time } \mathbf{t} \\
\begin{bmatrix} y^1_{1,1} & NA \\ y^1_{1,2} & NA \\ \vdots & \vdots \\ y^1_{N,n} & NA \\ NA & y^1_{1,1} \\ NA & y^1_{1,2} \\ \vdots & \vdots \\ NA & y^2_{N,n} \end{bmatrix}, &
\begin{bmatrix} 1 & NA \\ 1 & NA \\ \vdots & \vdots \\ 1 & NA \\ NA & 1 \\ NA & 1 \\ \vdots & \vdots \\ NA & 1 \end{bmatrix}, &
\begin{bmatrix} x_{1,1} & NA \\ x_{1,2} & NA \\ \vdots & \vdots \\ x_{N,n} & NA \\ NA & x_{1,1} \\ NA & x_{1,2} \\ \vdots & \vdots \\ NA & x_{N,n} \end{bmatrix}, &
\begin{bmatrix} t_{1,1} & NA \\ t_{1,2} & NA \\ \vdots & \vdots \\ t_{N,n} & NA \\ NA & t_{1,1} \\ NA & t_{1,2} \\ \vdots & \vdots \\ NA & t_{N,n} \end{bmatrix}
\end{array}$$

This ensures that INLA can fit two likelihoods and that INLA knows which covariates belong to which Likelihood.

- In order to fit the dependent random slope and random intercept, one has to use the 'iid2d' correlated random effect structure in INLA together with the 'copy' feature.
  An example of such a command in INLA is:

  ```
  f(Intercept, model="iid2d", n=2*N)+f(Time, time, copy="Intercept")
  ```

  Using this tells INLA that we shall be using a iid2d random effect structure (described in section 3.3.1) with a total of $2N$ random effects (2 for each individual, 1 random intercept + 1 random slope). The copy feature tells INLA that the 'Time' term is the second of these dependent random effects. One should also tell INLA which random effects are shared by the same object. If there are 2 measurements per person, the random effects for Intercept and Time should thus be written down as:

$$\begin{array}{cc}
\text{Random Intercept} & \text{Random Slope} \\
\begin{bmatrix} 1 & NA \\ 1 & NA \\ 2 & NA \\ \vdots & \vdots \\ N & NA \\ NA & 1 \\ NA & 1 \\ NA & 2 \\ \vdots & \vdots \\ NA & N \end{bmatrix}, &
\begin{bmatrix} 1 & NA \\ 1 & NA \\ 2 & NA \\ \vdots & \vdots \\ N & NA \\ NA & 1 \\ NA & 1 \\ NA & 2 \\ \vdots & \vdots \\ NA & N \end{bmatrix}
\end{array}$$

  This ensures that the random effects are unique per subject and in total we have $2 \cdot N$ different random effects.

The tricks presented in this section will be used to fit all subsequent models in this thesis. New R-INLA features will be discussed as they are used.
Model 0 can also be fit using the R-packages NLME, LMER and MCMCglmm. The details will not be presented here, but for those interested we refer to the open github repository where all code used during the thesis can be found `https://github.com/georgygomon/Thesis_open`.

4.2 Multivariate Joint Models

Having completed a base-line model which does not account for endogenous covariates we shall look at the first type of joint model which does take into account the association between the endogenous covariate and the outcome, Multivariate Joint Models, see section 2.3.1.

Definition 4.2 (Model 1A) *This Model, in which the association between the endogenous covariate y_1 and* *the outcome y_2 is modelled via residual errors, we shall be referring to as Model 1A.*

$$\begin{cases} y^1_{i,j} = \beta^1_0 + \beta^1_x \cdot x_i + \beta^1_t \cdot t_{i,j} + \epsilon^1_{i,j} \\ y^2_{i,j} = \beta^2_0 + \beta^2_x \cdot x_i + \beta^2_t \cdot t_{i,j} + \epsilon^2_{i,j} \end{cases} \quad \textit{with} \quad \begin{bmatrix} \epsilon^1_{i,1} \\ \epsilon^1_{i,2} \\ \epsilon^2_{i,1} \\ \epsilon^2_{i,2} \end{bmatrix} \sim \mathcal{N}_2 \left[\mathbf{0}, \begin{pmatrix} \sigma^2_{1,1} & \dots & \dots & \dots \\ \dots & \sigma^2_{1,2} & \dots & \dots \\ \dots & \dots & \sigma^2_{2,1} & \dots \\ \dots & \dots & \dots & \sigma^2_{2,2} \end{pmatrix} \right]$$

To fit this model using R-INLA an additional trick has to be used, as INLA does not allow for dependent residual errors.

- Since INLA does not allow for dependent residual errors, the residual errors should be modelled in INLA as random effects u rather than errors. Simultaneously, to get rid of the Gaussian noise of linear regression, one sets the Gaussian noise term fixed with a very high precision τ, in this way eliminating it.
  The resulting model is thus actually:

  $$\begin{cases} y^1_{i,j} = \beta^1_0 + \beta^1_x \cdot x_i + \beta^1_t \cdot t_{i,j} + u^1_{i,j} + \epsilon^1_{i,j} \\ y^2_{i,j} = \beta^2_0 + \beta^2_x \cdot x_i + \beta^2_t \cdot t_{i,j} + u^2_{i,j} + \epsilon^2_{i,j} \end{cases}$$

  with

  $$\begin{bmatrix} u^1_{i,1} \\ u^1_{i,2} \\ u^2_{i,1} \\ u^2_{i,2} \end{bmatrix} \sim \mathcal{N}_2 \left[\mathbf{0}, \begin{pmatrix} \sigma^2_{1,1} & \dots & \dots & \dots \\ \dots & \sigma^2_{1,2} & \dots & \dots \\ \dots & \dots & \sigma^2_{2,1} & \dots \\ \dots & \dots & \dots & \sigma^2_{2,2} \end{pmatrix} \right] \quad \text{and} \quad \begin{bmatrix} \boldsymbol{\epsilon}^1_i \\ \boldsymbol{\epsilon}^2_i \end{bmatrix} \sim \mathcal{N}_{2j}(\mathbf{0}, \tau \cdot \mathbf{I}_{2j})$$

  The random effects u have an *iid4d* distribution, as was discussed in section 3.3.1 and 4.1. Since the precision τ is set to be very high, the gaussian noise is thereby practically eliminated and thus this alternative representation in INLA corresponds very well to the actual Model 1A. However, within the R-INLA package one can not have an iid random effect with more than 5 components, and thus this implementation is limited to just 2 observations per subject ($3 \cdot 2 = 6 > 5$).

Model 1A can also be implemented in nlme (function gls) and MCMCglmm.

4.3 Joint Mixed Models

We shall now continue with Joint Mixed Models where the association between the endogenous covariate and the outcome is modelled via dependence of the random effects.

Definition 4.3 (Model 2A) *We start with a model in which the random intercept of both the endogenous covariate y_1 and the outcome y_2 are dependent. The association in time is given via the correlated residual errors. This model we shall refer to as model 2A.*

$$\begin{cases} y^1_{i,j} = (\beta^1_0 + u^1_{0,i}) + \beta^1_x \cdot x_i + \beta^1_t \cdot t_{i,j} + \epsilon^1_{i,j} \\ y^2_{i,j} = (\beta^2_0 + u^2_{0,i}) + \beta^2_x \cdot x_i + \beta^2_t \cdot t_{i,j} + \epsilon^2_{i,j} \end{cases}$$

with

$$\begin{bmatrix} u^1_{0,i} \\ u^2_{0,i} \end{bmatrix} \sim \mathcal{N}_2 \left[\mathbf{0}, \begin{pmatrix} \sigma^2_{1,0} & \sigma_{(1,2),0} \\ \sigma_{(2,1),0} & \sigma^2_{2,0} \end{pmatrix} \right]; \qquad \begin{bmatrix} \epsilon^1_{i,1} \\ \epsilon^1_{i,2} \\ \epsilon^2_{i,1} \\ \epsilon^2_{i,2} \end{bmatrix} \sim \mathcal{N}_2 \left[\mathbf{0}, \begin{pmatrix} \sigma^2_{1,1} & \dots & \dots & \dots \\ \dots & \sigma^2_{1,2} & \dots & \dots \\ \dots & \dots & \sigma^2_{2,1} & \dots \\ \dots & \dots & \dots & \sigma^2_{2,2} \end{pmatrix} \right]$$

To fit this model in INLA one simply combines the methods shown when discussing the models in Definitions 4.1 and 4.2. Model 2A can also be fit using nlme.

Definition 4.4 (Model 2B) *Next we consider a model in which both the association at baseline and in time between the endogenous covariate and the outcome are modelled via random effects. The residual errors are independent per measurement and play no role in the association between measurements or between y_1 and y_2.*

We shall call this Model 2B.

$$\begin{cases} y^1_{i,j} = (\beta^1_0 + u^1_{0,i}) + \beta^1_x \cdot x_i + \left(\beta^1_t + u^1_{t,i}\right) \cdot t_{i,j} + \epsilon^1_{i,j} \\ y^2_{i,j} = (\beta^2_0 + u^2_{0,i}) + \beta^2_x \cdot x_i + \left(\beta^2_t + u^2_{t,i}\right) \cdot t_{i,j} + \epsilon^2_{i,j} \end{cases}$$

with

$$\begin{bmatrix} u^1_{0,i} \\ u^2_{0,i} \\ u^1_{t,i} \\ u^2_{t,i} \end{bmatrix} \sim \mathcal{N}_4 \left[\mathbf{0}, \begin{pmatrix} \sigma^2_{1,0} & \sigma_{(1,0),(2,0)} & \sigma_{(1,0),(1,t)} & \sigma_{(1,0),(2,t)} \\ \sigma_{(2,0),(1,0)} & \sigma^2_{2,0} & \sigma_{(2,0),(1,t)} & \sigma_{(2,0),(2,t)} \\ \sigma_{(1,t),(1,0)} & \sigma_{(1,t),(2,0)} & \sigma^2_{1,t} & \sigma_{(1,t),(2,t)} \\ \sigma_{(2,t),(1,0)} & \sigma_{(2,t),(2,0)} & \sigma_{(2,t),(1,t)} & \sigma^2_{2,t} \end{pmatrix} \right] \quad ; \quad \begin{bmatrix} \boldsymbol{\epsilon}^1_i \\ \boldsymbol{\epsilon}^2_i \end{bmatrix} \sim \mathcal{N}_{2j}(\mathbf{0}, \mathbf{I}_{2j})$$

This model can be fit in INLA using the 'iid4d' random effects (see section 3.3.1) and the tricks shown in definition 4.1. Furthermore, this model can also be fit using NLME, LMER and MCMCglmm.

Definition 4.5 (Model 2B1) *Model 2B1 is a small modification from model 2B and the 2 models are nested. The only difference is that in model 2B1 there is no dependence between the random slope and random intercept in either y_1 or y_2. Thus, the covariance matrix of the random effects becomes:*

$$\begin{bmatrix} u^1_{0,i} \\ u^2_{0,i} \\ u^1_{t,i} \\ u^2_{t,i} \end{bmatrix} \sim \mathcal{N}_4 \left[\mathbf{0}, \begin{pmatrix} \sigma^2_{1,0} & \sigma_{(1,0),(2,0)} & 0 & 0 \\ \sigma_{(2,0),(1,0)} & \sigma^2_{2,0} & 0 & 0 \\ 0 & 0 & \sigma^2_{1,t} & \sigma_{(1,t),(2,t)} \\ 0 & 0 & \sigma_{(2,t),(1,t)} & \sigma^2_{2,t} \end{pmatrix} \right]$$

This model was introduced to test, based on the goodness of fit measures, how similar these 2 nested models would perform on simulated data.

4.4 Joint Mixed Models with Scaled Linear Predictor

The last type of models we shall be considering in this thesis are models in which the linear predictor for the endogenous covariate, y_1, is copied with a scaling factor γ into the linear predictor of the outcome y_1.

Definition 4.6 (Model 3A) *In this model the entire linear predictor of the endogenous covariate y_1 is copied into the linear predictor of the outcome y_2 with scaling factor gamma. In both linear predictors we have dependent random intercept and random slope terms. The model will be referred to as Model 3A and is mathematically given by:*

$$\begin{cases} m_{ij} = (\beta^1_0 + u^1_{0,i}) + \beta^1_x \cdot x_i + \left(\beta^1_t + u^1_{t,i}\right) \cdot t_{i,j} \\ y^1_{i,j} = m_{ij} + \epsilon^1_{i,j} \\ y^2_{i,j} = \gamma \cdot m_{ij} + (\beta^2_0 + u^2_{0,i}) + \beta^2_x \cdot x_i + \left(\beta^2_t + u^2_{t,i}\right) \cdot t_{i,j} + \epsilon^2_{i,j} \end{cases}$$

with

$$\begin{bmatrix} u^1_{0,i} \\ u^1_{t,i} \end{bmatrix} \sim \mathcal{N}_2 \left[\mathbf{0}, \begin{pmatrix} \sigma^2_{1,0} & \sigma_{1,(0,t)} \\ \sigma_{1,(t,0)} & \sigma^2_{1,t} \end{pmatrix} \right] ; \quad \begin{bmatrix} u^2_{0,i} \\ u^2_{t,i} \end{bmatrix} \sim \mathcal{N}_2 \left[\mathbf{0}, \begin{pmatrix} \sigma^2_{2,0} & \sigma_{2,(0,t)} \\ \sigma_{2,(t,0)} & \sigma^2_{2,t} \end{pmatrix} \right] ; \quad \begin{bmatrix} \boldsymbol{\epsilon}^1_i \\ \boldsymbol{\epsilon}^2_i \end{bmatrix} \sim \mathcal{N}_{2j}(\mathbf{0}, \mathbf{I}_{2j})$$

In order to implement this model within R-INLA a few tricks need to be used not previously discussed:

- INLA allows for random effects to be copied with a scaling factor γ into a different likelihood. For this one uses the following syntax:

  ```
  f(Intercept1)+ f(Intercept12, copy="Intercept1", hyper = list(beta = list(fixed=FALSE)))
  ```

  Hereby we indicate that we want to copy the element Intercept1 into a different likelihood with a non-fixed scaling factor *gamma*.

- To ensure that all elements being copied use the same scaling factor γ, the following syntax is used:

  ```
  f(x1, x)+
    f(x12, x, copy="x1", same.as = 'Intercept12', hyper = list(beta = list(fixed=FALSE)))
  ```

  Hereby we ensure that the element (x1) is copied with a scaling factor γ that is the same as was used when copying and scaling 'Intercept12'.

- A problem within INLA is that only random effects can be copied and scaled in this way. Thus, the only way to copy and scale fixed effects is to turn them into random effects with 2 levels, one for y_1 and one for y_2. As example, we would have a random effect for the Intercept $\beta_{0,k} \sim \mathcal{N}(0, \sigma_0^2), \quad k = 1, 2$, with 2 levels, one for y_1 and one for y_2, equal for all subjects. We are then not interested in the variance σ_0^2 of this random effect but instead in the realisation of the random effect for both the endogenous covariate y_1 and the outcome y_2.
  In this way all of the fixed effects are written down as random effects and copied with the same scaling parameter. Particularly within the Bayesian framework the difference between fixed and random effects is more subtle than in a frequentist approach, as both fixed and random effects have parameters that are random variables.
- Random effects are implemented as was discussed in definition 4.1, and copied in the same way as are fixed effects.

Definition 4.7 (Model 3A1) *Model 3A1 is a small modification from model 3A and the 2 models are nested. The only difference is that in model 3A1 there is no dependence between the random slope and random intercept in y_1. Thus, the covariance matrix of the random effects for the endogenous covariate y_1 becomes:*

$$\begin{bmatrix} u_{0,i}^1 \\ u_{t,i}^1 \end{bmatrix} \sim \mathcal{N}_2 \left[\mathbf{0}, \begin{pmatrix} \sigma_{1,0}^2 & 0 \\ 0 & \sigma_{1,t}^2 \end{pmatrix} \right]$$

This model was introduced to test, based on the goodness of fit measures, how similar these 2 nested models would perform on simulated data.

Definition 4.8 (Model 3B) *Model 3B was inspired by [Guo and Carlin, 2004], in which many models with scaled linear predictors were discussed. In Model 3B the random intercept and random slope are copied and scaled with different scaling parameters γ_1 and γ_2.*

$$\begin{cases} y_{i,j}^1 = (\beta_0^1 + u_{0,i}^1) + \beta_x^1 \cdot x_i + \left(\beta_t^1 + u_{t,i}^1\right) \cdot t_{i,j} + \epsilon_{i,j}^1 \\ y_{i,j}^2 = \gamma_1 \cdot u_{0,i}^1 + \gamma_2 \cdot u_{t,i}^1 + (\beta_0^2 + u_{0,i}^2) + \beta_x^2 \cdot x_i + \left(\beta_t^2 + u_{t,i}^2\right) \cdot t_{i,j} + \epsilon_{i,j}^2 \end{cases}$$

with

$$u_{0,i}^1 \sim \mathcal{N}(0, \sigma_{1,0}^2); \quad u_{t,i}^1 \sim \mathcal{N}(0, \sigma_{1,t}^2); \quad \begin{bmatrix} u_{0,i}^2 \\ u_{t,i}^2 \end{bmatrix} \sim \mathcal{N}_2 \left[\mathbf{0}, \begin{pmatrix} \sigma_{2,0}^2 & \sigma_{2,(0,t)} \\ \sigma_{2,(t,0)} & \sigma_{2,t}^2 \end{pmatrix} \right]; \quad \begin{bmatrix} \boldsymbol{\epsilon}_i^1 \\ \boldsymbol{\epsilon}_i^2 \end{bmatrix} \sim \mathcal{N}_{2j}(\mathbf{0}, \mathbf{I}_{2j})$$

For the models with scaled linear predictor it is important to note that they can be rewritten into a form where there is no dependence between y_1 and y_2. If we take Model 3A as example, we can write the linear predictor for y_2 as:

$$\begin{aligned} y_{i,j}^2 &= \gamma \cdot m_{ij} + (\beta_0^2 + u_{0,i}^2) + \beta_x^2 \cdot x_i + \left(\beta_t^2 + u_{t,i}^2\right) \cdot t_{i,j} + \epsilon_{i,j}^2 = \\ &= (\gamma\beta_0^1 + \beta_0^2) + (\gamma\beta_x^1 + \beta_x^2)x_i + \left(\gamma\beta_t^1 + \beta_t^2\right) \cdot t_{i,j} + (\gamma u_{0,i}^1 + u_{0,i}^2) + (\gamma u_{t,i}^1 + u_{t,i}^2)t_{i,j} + \epsilon_{i,j}^2 = \\ &= \beta_0^{2'} + \beta_x^{2'} x_i + \beta_t^{2'} t_{i,j} + u_{0,i}^{2'} + u_{t,i}^{2'} \cdot t_{i,j} + \epsilon_{i,j}^2 \end{aligned}$$

with $\beta_0^{2'} = \gamma\beta_0^1 + \beta_0^2$, $\beta_x^{2'} = \gamma\beta_x^1 + \beta_x^2$ and $\beta_t^{2'} = \gamma\beta_t^1 + \beta_t^2$.
For the random effects, they now have distribution:

$$\begin{bmatrix} u_{0,i}^{2'} \\ u_{t,i}^{2'} \end{bmatrix} \sim \mathcal{N}_2 \left[\mathbf{0}, \begin{pmatrix} \gamma^2\sigma_{1,0}^2 + \sigma_{2,0}^2 & \gamma^2\sigma_{1,(0,t)} + \sigma_{2,(0,t)} \\ \gamma^2\sigma_{1,(t,0)} + \sigma_{2,(t,0)} & \gamma^2\sigma_{1,t}^2 + \sigma_{2,t}^2 \end{pmatrix} \right]$$

Thus, although the models with the scaled linear predictor do seem to be very different from the independent joint models, the scaled linear predictor model can be rewritten in the same form as Model 0, see definition 4.1.
Further note that none of the models with scaled linear predictor can be fit using the packages nlme, lmer or MCMCglmm.

4.5 Summary of Models introduced

A summary of all the Models introduced in the previous sections is given in table 1. Note that all models have the same fixed part, consisting of:

$$\begin{cases} y_{i,j}^1 = \beta_0^1 + \beta_x^1 \cdot x_{i,j} + \beta_t^1 t_{i,j} \\ y_{i,j}^2 = \beta_0^2 + \beta_x^2 \cdot x_{i,j} + \beta_t^2 t_{i,j} \end{cases}$$

The subsequent random effects, errors and scaling elements unique to each model are given in table 1.

Model	Random effects	Errors	Scaling
0	$\begin{bmatrix} u^1_{0,i} \\ u^1_{t,i} \end{bmatrix} \sim \mathcal{N}_2 \left[\mathbf{0}, \begin{pmatrix} \sigma^2_{1,0} & \sigma_{1,(0,t)} \\ \sigma_{1,(t,0)} & \sigma^2_{1,t} \end{pmatrix} \right]$ $\begin{bmatrix} u^2_{0,i} \\ u^2_{t,i} \end{bmatrix} \sim \mathcal{N}_2 \left[\mathbf{0}, \begin{pmatrix} \sigma^2_{2,0} & \sigma_{2,(0,t)} \\ \sigma_{2,(t,0)} & \sigma^2_{2,t} \end{pmatrix} \right]$	$\begin{bmatrix} \boldsymbol{\epsilon}^1_i \\ \boldsymbol{\epsilon}^2_i \end{bmatrix} \sim \mathcal{N}_{2j}(\mathbf{0}, \mathbf{I}_{2j})$	✗
1A	✗	$\begin{bmatrix} \epsilon^1_{i,1} \\ \epsilon^1_{i,2} \\ \epsilon^2_{i,1} \\ \epsilon^2_{i,2} \end{bmatrix} \sim \mathcal{N}_4 \left[\mathbf{0}, \begin{pmatrix} \sigma^2_{1,1} & ... & ... & ... \\ ... & \sigma^2_{1,2} & ... & ... \\ ... & ... & \sigma^2_{2,1} & ... \\ ... & ... & ... & \sigma^2_{2,2} \end{pmatrix} \right]$	✗
2A	$\begin{bmatrix} u^1_{0,i} \\ u^2_{0,i} \end{bmatrix} \sim \mathcal{N}_4 \left[\mathbf{0}, \begin{pmatrix} \sigma^2_{1,0} & \sigma_{(1,2),0} \\ \sigma_{(2,1),0} & \sigma^2_{2,0} \end{pmatrix} \right]$	$\begin{bmatrix} \epsilon^1_{i,1} \\ \epsilon^1_{i,2} \\ \epsilon^2_{i,1} \\ \epsilon^2_{i,2} \end{bmatrix} \sim \mathcal{N}_2 \left[\mathbf{0}, \begin{pmatrix} \sigma^2_{1,1} & ... & ... & ... \\ ... & \sigma^2_{1,2} & ... & ... \\ ... & ... & \sigma^2_{2,1} & ... \\ ... & ... & ... & \sigma^2_{2,2} \end{pmatrix} \right]$	✗
2B	$\begin{bmatrix} u^1_{0,i} \\ u^2_{0,i} \\ u^1_{t,i} \\ u^2_{t,i} \end{bmatrix} \sim \mathcal{N}_4 \left[\mathbf{0}, \begin{pmatrix} \sigma^2_{1,0} & ... & ... & ... \\ ... & \sigma^2_{2,0} & ... & ... \\ ... & ... & \sigma^2_{1,t} & ... \\ ... & ... & ... & \sigma^2_{2,t} \end{pmatrix} \right]$	$\begin{bmatrix} \boldsymbol{\epsilon}^1_i \\ \boldsymbol{\epsilon}^2_i \end{bmatrix} \sim \mathcal{N}_{2j}(\mathbf{0}, \mathbf{I}_{2j})$	✗
2B1	$\begin{bmatrix} u^1_{0,i} \\ u^2_{0,i} \\ u^1_{t,i} \\ u^2_{t,i} \end{bmatrix} \sim \mathcal{N}_4 \left[\mathbf{0}, \begin{pmatrix} \sigma^2_{1,0} & ... & 0 & 0 \\ ... & \sigma^2_{2,0} & 0 & 0 \\ 0 & 0 & \sigma^2_{1,t} & ... \\ 0 & 0 & ... & \sigma^2_{2,t} \end{pmatrix} \right]$	$\begin{bmatrix} \boldsymbol{\epsilon}^1_i \\ \boldsymbol{\epsilon}^2_i \end{bmatrix} \sim \mathcal{N}_{2j}(\mathbf{0}, \mathbf{I}_{2j})$	✗
3A	$\begin{bmatrix} u^1_{0,i} \\ u^1_{t,i} \end{bmatrix} \sim \mathcal{N}_2 \left[\mathbf{0}, \begin{pmatrix} \sigma^2_{1,0} & \sigma_{1,(0,t)} \\ \sigma_{1,(t,0)} & \sigma^2_{1,t} \end{pmatrix} \right]$ $\begin{bmatrix} u^2_{0,i} \\ u^2_{t,i} \end{bmatrix} \sim \mathcal{N}_2 \left[\mathbf{0}, \begin{pmatrix} \sigma^2_{2,0} & \sigma_{2,(0,t)} \\ \sigma_{2,(t,0)} & \sigma^2_{2,t} \end{pmatrix} \right]$	$\begin{bmatrix} \boldsymbol{\epsilon}^1_i \\ \boldsymbol{\epsilon}^2_i \end{bmatrix} \sim \mathcal{N}_{2j}(\mathbf{0}, \mathbf{I}_{2j})$	$y^1_{i,j} = m_{ij} + \epsilon^1_{i,j}$ $y^2_{i,j} = \gamma m_{ij} + ...$
3A1	$\begin{bmatrix} u^1_{0,i} \\ u^1_{t,i} \end{bmatrix} \sim \mathcal{N}_2 \left[\mathbf{0}, \begin{pmatrix} \sigma^2_{1,0} & 0 \\ 0 & \sigma^2_{1,t} \end{pmatrix} \right]$ $\begin{bmatrix} u^2_{0,i} \\ u^2_{t,i} \end{bmatrix} \sim \mathcal{N}_2 \left[\mathbf{0}, \begin{pmatrix} \sigma^2_{2,0} & \sigma_{2,(0,t)} \\ \sigma_{2,(t,0)} & \sigma^2_{2,t} \end{pmatrix} \right]$	$\begin{bmatrix} \boldsymbol{\epsilon}^1_i \\ \boldsymbol{\epsilon}^2_i \end{bmatrix} \sim \mathcal{N}_{2j}(\mathbf{0}, \mathbf{I}_{2j})$	$y^1_{i,j} = m_{ij} + \epsilon^1_{i,j}$ $y^2_{i,j} = \gamma m_{ij} + ...$
3B	$\begin{bmatrix} u^1_{0,i} \\ u^1_{t,i} \end{bmatrix} \sim \mathcal{N}_2 \left[\mathbf{0}, \begin{pmatrix} \sigma^2_{1,0} & 0 \\ 0 & \sigma^2_{1,t} \end{pmatrix} \right]$ $\begin{bmatrix} u^2_{0,i} \\ u^2_{t,i} \end{bmatrix} \sim \mathcal{N}_2 \left[\mathbf{0}, \begin{pmatrix} \sigma^2_{2,0} & \sigma_{2,(0,t)} \\ \sigma_{2,(t,0)} & \sigma^2_{2,t} \end{pmatrix} \right]$	$\begin{bmatrix} \boldsymbol{\epsilon}^1_i \\ \boldsymbol{\epsilon}^2_i \end{bmatrix} \sim \mathcal{N}_{2j}(\mathbf{0}, \mathbf{I}_{2j})$	$y^2_{i,j} = \gamma_1 u^1_{0,i} +$ $+ \gamma_2 u^1_{t,i}$

Table 1: Summary of all models introduced in the previous sections

4.6 Joint Models and their implementation in different R packages

We have attempted to fit all models discussed so far using the R packages R-INLA, nlme, lmer and MCM-Cglmm. The packages R-INLA and MCMCglmm are Bayesian packages for Mixed Models, with R-INLA using the INLA framework for inference while MCMCglmm uses MCMC sampling. Both nlme and lmer are frequentist methods for solving Mixed Models. An overview of the different models and the packages that are able to fit them is given in table 2.

		R-INLA	NLME	LMER	MCMCglmm
Independent Mixed Models	Only Random Effects	✓	✓	✓	✓
Multivariate Joint Model	Correlated Residual Errors	✓	✓	✗	✓
Joint Mixed Models	Random effect + correlated residual errors	✓	✓	✗	✗
	Only random effects	✓	✓	✓	✓
Joint Mixed Models with scaled linear predictor	Only Random Effects	✓	✗	✗	✗

Table 2: Tabel indicating which R-packages can fit which joint models.

5 Simulation Study

In order to test the implementation of the different joint models in R we used simulated data. To construct the simulated data a total of $i = 1, ..., N$ patients were simulated with $j = 1, ..., n$ measurements per patient. Per measurement a variable x was randomly sampled from the $\mathcal{N}(0, 1)$ distribution, and the progression over time was modelled to be linear. In order to make the data unbalanced at every measurement time there was a probability p_1 that the exogenous covariate y_1 was measured and a probability p_2 that the outcome y_2 was measured. An example of such a data-set is shown in table 3.

Table 3: Table showing the head of simulated dataset

id	x	time	y_1 observed	y_2 observed
1	-0.67	0	1	0
1	-1.18	1	0	1
1	0.87	2	1	0
2	0.11	0	1	1
2	-1.36	1	0	1
2	-0.08	2	0	0

The endogenous covariate and the outcome were simulated according to the different models, with the addition of an extra error sampled from a $\mathcal{N}(0, 0.5)$ distribution.

5.1 Results of Simulation Study

Having been able to fit all above mentioned models using R-INLA we did a simulation study to compare the goodness of fit of these models. We proceeded as follows:

- We simulated data as described in section 3. The endogenous covariate y_1 and the outcome y_2 we simulated according to each model. Thus, in total we simulated 8 data-sets.
- We then fitted each data-set with each of the models. We hereby calculated all goodness of fit measures available (see section 3.2), which include the Marginal Likelihood, DIC, WAIC, CPO, PIT and the three different types of MSE (see section 3.2.5). To test the correctness of the Marginal Likelihood and the DIC, we also approximated the DIC via $-2 \log MLIK$, as this approximation is generally correct, [].
- For each model we generated 5 datasets and took the average goodness of fit measures over these 5 datasets. Hereby we thus performed a 5-fold Cross-validation. Since the data is generated we simply used this approach instead of generating a big dataset and splitting it into 5 pieces.
- In total we thus fitted 8 (Datasets generated according to each Model) $\times 5$ (Each dataset generated 5 times for 5-fold CV) $\times 8$ (Each dataset generated is fit on each model = 320 models.

- The parameters used are: $N = 75$, $n = 6$, $p_1 = 0.7$, $p_2 = 0.7$, $CV = 5$, $\boldsymbol{\beta} = (\beta_0^1, \beta_x^1, \beta_t^1, \beta_0^2, \beta_x^2, \beta_t^2) = (2.4, 2.5, 3, 1.5, 2.5)$.
  To calculate the MSE on subsequent measurements of subjects included in fitting the model (to determine how well the model fitted the random effects), an additional $n =$ measurements per subject were sampled (but not used in the fitting process).
  To calculate the MSE on test subjects (to determine the marginal results) an additional $N = 25$ subjects with $n = 10$ measurements were sampled but not included in the fitting process.
- TODO!!!! ALL MODEL PARAMETERS!!

As models 1A and 2A can not handle more than 2 measurements per subject (because the residual error structure can not account for more than 2 measurements per subject), these models were not used in the simulations.

The results of this simulation studies can be seen in figures 1 and 2.
TODO!!!! WRITE DOWN SIMULATION STUDIES NICELY!!!!

	Model_0	Model_2A	Model_2C1	Model_2C2	Model_3A1	Model_3A2	Model_3B1
MLIK	-1169.389	-1932.045	-1178.876	-1207.703	-1219.027	-1200.063	-1197.392
DIC_approx	2338.778	3864.090	2357.751	2415.405	2438.055	2400.127	2394.784
DIC	1209.752	3519.989	1209.602	1215.059	1209.653	1210.469	1210.495
WAIC	1195.581	3531.846	1192.782	1201.509	1193.481	1195.920	1194.475
PIT	0.025	0.088	0.025	0.025	0.025	0.025	0.027
CPO	677.870	1774.285	680.786	679.700	679.311	678.085	679.858
MSE_train	0.331	10.789	0.334	0.350	0.330	0.337	0.328
MSE_same	1.571	98.078	1.592	1.686	1.561	1.610	1.538
MSE_others	127.771	128.026	127.761	127.771	128.252	128.227	127.771

	Model_0	Model_2A	Model_2C1	Model_2C2	Model_3A1	Model_3A2	Model_3B1
	-847.507	-793.282	-841.268	-847.495	-857.238	-877.256	-831.768
	1695.014	1586.565	1682.537	1694.990	1714.475	1754.512	1663.536
	1112.649	1072.814	1111.871	1111.300	1083.564	1111.989	1083.493
	1104.172	1078.681	1104.977	1103.328	1081.711	1103.509	1081.663
	0.027	0.028	0.027	0.030	0.026	0.027	0.026
	588.087	549.665	586.576	584.822	565.006	587.507	564.979
	0.212	0.232	0.213	0.213	0.221	0.212	0.221
	0.531	0.315	0.525	0.513	0.439	0.535	0.439
	2.668	2.669	2.668	2.668	2.667	2.667	2.669

	Model_0	Model_2A	Model_2C1	Model_2C2	Model_3A1	Model_3A2	Model_3B1
MLIK	-1178.191	-1923.709	-1166.211	-1204.087	-1222.966	-1207.838	-1201.407
DIC_approx	2356.383	3847.418	2332.421	2408.174	2445.931	2415.675	2402.815
DIC	1215.790	3534.725	1213.568	1219.995	1214.175	1218.136	1213.943
WAIC	1199.795	3544.798	1198.561	1205.480	1198.340	1202.246	1197.411
PIT	0.026	0.084	0.026	0.025	0.027	0.025	0.026
CPO	685.908	1779.544	683.051	685.469	684.974	687.501	684.313
MSE_train	0.330	11.147	0.327	0.342	0.329	0.338	0.328
MSE_same	1.550	100.414	1.513	1.604	1.537	1.594	1.526
MSE_others	118.796	119.150	118.842	118.859	118.753	118.550	118.786

	Model_0	Model_2A	Model_2C1	Model_2C2	Model_3A1	Model_3A2	Model_3B1
	-1166.631	-1919.015	-1165.260	-1202.957	-1218.051	-1196.474	-1198.868
	2333.262	3838.031	2330.520	2405.915	2436.101	2392.949	2397.735
	1211.594	3499.312	1209.069	1293.185	1209.890	1213.642	1209.276
	1196.606	3509.455	1193.022	1282.842	1193.234	1198.930	1192.738
	0.025	0.076	0.026	0.030	0.028	0.025	0.028
	678.972	1762.525	679.304	709.608	680.182	680.118	679.733
	0.362	10.696	0.342	0.430	0.346	0.360	0.352
	1.706	95.984	1.583	2.008	1.618	1.691	1.657
	135.476	135.895	135.486	135.490	135.431	135.173	135.464

Figure 1: Results of the simulation study. In red one can see the models under which the data is simulated.

	Model_0	Model_2A	Model_2C1	Model_2C2	Model_3A1	Model_3A2	Model_3B1
MLIK	-1227.854	-1591.308	-1183.693	-1278.061	-1263.796	-1204.950	-1227.723
DIC_approx	2455.708	3182.616	2367.385	2556.122	2527.592	2409.901	2455.446
DIC	1216.514	2949.277	1205.455	1268.172	1213.095	1205.181	1207.784
WAIC	1199.852	2958.381	1189.583	1261.789	1195.514	1190.450	1191.940
PIT	0.024	0.073	0.027	0.025	0.027	0.027	0.027
CPO	688.405	1483.705	674.549	701.766	685.743	672.278	675.866
MSE_train	0.580	13.706	0.329	0.551	0.652	0.325	0.559
MSE_same	3.152	123.240	1.523	2.835	3.631	1.496	2.956
MSE_others	171.291	138.898	171.184	171.171	171.818	171.468	171.295

	Model_0	Model_2A	Model_2C1	Model_2C2	Model_3A1	Model_3A2	Model_3B1
	-1204.601	-2011.635	-1182.344	-1224.154	-1252.239	-1190.376	-1222.585
	2409.201	4023.271	2364.688	2448.308	2504.479	2380.752	2445.171
	1208.557	3700.138	1202.307	1301.895	1208.780	1199.803	1206.367
	1194.134	3712.664	1185.706	1290.747	1191.563	1185.574	1189.725
	0.024	0.075	0.027	0.025	0.025	0.025	0.027
	676.914	1862.755	672.290	709.253	681.137	666.265	675.740
	0.493	16.634	0.329	0.380	0.590	0.327	0.540
	2.584	148.685	1.510	1.762	3.196	1.503	2.826
	232.776	232.939	232.757	236.125	233.238	233.161	232.816

	Model_0	Model_2A	Model_2C1	Model_2C2	Model_3A1	Model_3A2	Model_3B1
MLIK	-1232.779	-2041.617	-1182.779	-1270.669	-1254.730	-1263.022	-1225.741
DIC_approx	2465.559	4083.235	2365.558	2541.338	2509.459	2526.045	2451.482
DIC	1217.195	3786.446	1209.815	1328.218	1208.223	1217.962	1208.337
WAIC	1201.008	3798.907	1195.072	1320.722	1192.515	1201.535	1193.633
PIT	0.022	0.085	0.026	0.024	0.025	0.024	0.026
CPO	690.169	1905.339	677.676	731.143	678.944	690.400	676.885
MSE_train	0.559	19.506	0.347	0.450	0.548	0.571	0.576
MSE_same	3.066	174.779	1.648	2.169	2.991	3.145	3.138
MSE_others	177.512	179.778	176.923	177.194	177.665	178.091	177.510

Figure 2: Results of the simulation study. In red one can see the models under which the data is simulated.

References

[Guo and Carlin, 2004] Guo, X. and Carlin, B. P. (2004). Separate and Joint Modeling of Longitudinal and Event Time Data Using Standard Computer Packages. *American Statistician*, 58(1):16–24.

[P.J. Diggle, 2016] P.J. Diggle, P. H. (2016). Analysis of Longitudinal Data (Book). *Book*, (April):5–24.

[Rue et al., 2009] Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 71(2):319–392.

[van Niekerk et al., 2021] van Niekerk, J., Bakka, H., and Rue, H. (2021). Competing risks joint models using R-INLA. *Statistical modelling*, 21(1-2):56–71.