Linear mixed models with improper priors and flexible distributional assumptions for longitudinal and survival data

Francisco Javier Rubio
Joint work with Prof. Mark F. J. Steel

University of Warwick
Department of Statistics

University of Warwick, 2015

Table of contents

The standard model structure

- Consider the hierarchical linear mixed model (LMM):

$$y_{ij} = \mathbf{x}_{ij}^{\top}\beta + \mathbf{z}_{ij}^{\top}\mathbf{u}_i + \varepsilon_{ij}, \tag{1}$$

where $\mathbf{y} = \{y_{ij}\}$ denotes the $n \times 1$ vector of response variables for subject i at time t_{ij}, $j = 1, \ldots, n_i$ denotes the number of repeated measurements for subject i, $i = 1, \ldots, r$ denotes the number of subjects, β is a $p \times 1$ vector of *fixed effects*, $\mathbf{u}_i$ are $q \times 1$ mutually independent random vectors and ε_{ij} are *i.i.d.* errors.

- In matrix notation we can write model (1) as follows:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \varepsilon,$$

where $\mathbf{X}$ and $\mathbf{Z}$ denote the known design matrices of dimension $n \times p$ and $n \times q$, respectively, ε the $n \times 1$ vector of errors, and $\mathbf{u} = (\mathbf{u}_1^{\top}, \ldots, \mathbf{u}_r^{\top})^{\top}$.

The standard model structure

- Consider the hierarchical linear mixed model (LMM):

$$y_{ij} = \mathbf{x}_{ij}^{\top}\beta + \mathbf{z}_{ij}^{\top}\mathbf{u}_i + \varepsilon_{ij}, \tag{1}$$

where $\mathbf{y} = \{y_{ij}\}$ denotes the $n \times 1$ vector of response variables for subject i at time t_{ij}, $j = 1, \ldots, n_i$ denotes the number of repeated measurements for subject i, $i = 1, \ldots, r$ denotes the number of subjects, β is a $p \times 1$ vector of *fixed effects*, $\mathbf{u}_i$ are $q \times 1$ mutually independent random vectors and ε_{ij} are *i.i.d.* errors.

- In matrix notation we can write model (1) as follows:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \varepsilon,$$

where $\mathbf{X}$ and $\mathbf{Z}$ denote the known design matrices of dimension $n \times p$ and $n \times q$, respectively, ε the $n \times 1$ vector of errors, and $\mathbf{u} = (\mathbf{u}_1^{\top}, \ldots, \mathbf{u}_r^{\top})^{\top}$.

These models are used in different areas under different names and notation.

- In Bayesian analysis of variance, the use of these models goes back to Tiao and Tan (1965).

These models are used in different areas under different names and notation.

- In Bayesian analysis of variance, the use of these models goes back to Tiao and Tan (1965). **LMM** for longitudinal data.

These models are used in different areas under different names and notation.

- In Bayesian analysis of variance, the use of these models goes back to Tiao and Tan (1965). **LMM** for longitudinal data.
- In survival analysis, the logarithm of the survival times $\mathbf{T} = (T_1, \ldots, T_n)$ are modelled using a LMM. This is

$$\log(\mathbf{T}) = \mathbf{X}\beta + \mathbf{Zu} + \varepsilon, \tag{2}$$

Model (2) is often referred to as a Mixed Effects Accelerated Failure Time model (**MEAFT**).

These models are used in different areas under different names and notation.

- In Bayesian analysis of variance, the use of these models goes back to Tiao and Tan (1965). **LMM** for longitudinal data.
- In survival analysis, the logarithm of the survival times $\mathbf{T} = (T_1, \ldots, T_n)$ are modelled using a LMM. This is

$$\log(\mathbf{T}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \tag{2}$$

  Model (2) is often referred to as a Mixed Effects Accelerated Failure Time model (**MEAFT**).
- In the context of econometrics, the logarithm of the output (or the negative of the logarithm of the cost) of q firms are modeled using (1) with certain restrictions on the regression coefficients$\boldsymbol{\beta}$ as well as the random effects $\mathbf{u}$. Under these additional conditions, the resulting model is referred to as the **Linear Stochastic Frontier Model** (see *e.g.* Fernández et al., 1997).

Distributional assumptions

- The distribution of the errors ε_{ij} and the random effects u_{ij} is typically assumed to be normal.

Distributional assumptions

- The distribution of the errors ε_{ij} and the random effects u_{ij} is typically assumed to be normal.
- There is a common agreement that the correct specification of the distribution of the errors is relevant (e.g. heavy tails).

Distributional assumptions

- The distribution of the errors ε_{ij} and the random effects u_{ij} is typically assumed to be normal.
- There is a common agreement that the correct specification of the distribution of the errors is relevant (e.g. heavy tails).
- There is a lot of debate on whether the correct specification of the distribution of the random effects is relevant or not.

Distributional assumptions

- The distribution of the errors ε_{ij} and the random effects u_{ij} is typically assumed to be normal.
- There is a common agreement that the correct specification of the distribution of the errors is relevant (e.g. heavy tails).
- There is a lot of debate on whether the correct specification of the distribution of the random effects is relevant or not. Typically, the answer depends on the aims (Zhang and Davidian, 2001).

Distributional assumptions

- The distribution of the errors ε_{ij} and the random effects u_{ij} is typically assumed to be normal.
- There is a common agreement that the correct specification of the distribution of the errors is relevant (e.g. heavy tails).
- There is a lot of debate on whether the correct specification of the distribution of the random effects is relevant or not. Typically, the answer depends on the aims (Zhang and Davidian, 2001).
- Some flexible extensions have been proposed:

Distributional assumptions

- The distribution of the errors ε_{ij} and the random effects u_{ij} is typically assumed to be normal.
- There is a common agreement that the correct specification of the distribution of the errors is relevant (e.g. heavy tails).
- There is a lot of debate on whether the correct specification of the distribution of the random effects is relevant or not. Typically, the answer depends on the aims (Zhang and Davidian, 2001).
- Some flexible extensions have been proposed:
  1. Zhang and Davidian (2001) suppose that the random effects are distributed according to a finite mixture of normals, and normal errors.

Distributional assumptions

- The distribution of the errors ε_{ij} and the random effects u_{ij} is typically assumed to be normal.
- There is a common agreement that the correct specification of the distribution of the errors is relevant (e.g. heavy tails).
- There is a lot of debate on whether the correct specification of the distribution of the random effects is relevant or not. Typically, the answer depends on the aims (Zhang and Davidian, 2001).
- Some flexible extensions have been proposed:
  1. Zhang and Davidian (2001) suppose that the random effects are distributed according to a finite mixture of normals, and normal errors.
  2. Komárek and Lesaffre (2007) suppose that the errors are distributed according to finite mixture of normals, and normal random effects.

Distributional assumptions

- The distribution of the errors ε_{ij} and the random effects u_{ij} is typically assumed to be normal.
- There is a common agreement that the correct specification of the distribution of the errors is relevant (e.g. heavy tails).
- There is a lot of debate on whether the correct specification of the distribution of the random effects is relevant or not. Typically, the answer depends on the aims (Zhang and Davidian, 2001).
- Some flexible extensions have been proposed:
  1. Zhang and Davidian (2001) suppose that the random effects are distributed according to a finite mixture of normals, and normal errors.
  2. Komárek and Lesaffre (2007) suppose that the errors are distributed according to finite mixture of normals, and normal random effects.
  3. Lachos et al. (2010) suppose that the errors and the random effects are jointly distributed as scale mixtures of multivariate skew normals.

Non-informative priors

- “Noninformative” or “Benchmark” priors are of interest in Bayesian analysis when the prior information is vague.

Non-informative priors

- “Noninformative” or “Benchmark” priors are of interest in Bayesian analysis when the prior information is vague.
- These priors typically produce posteriors with good frequentist properties.

Non-informative priors

- “Noninformative” or “Benchmark” priors are of interest in Bayesian analysis when the prior information is vague.
- These priors typically produce posteriors with good frequentist properties.
- They are typically improper:

Non-informative priors

- “Noninformative” or “Benchmark” priors are of interest in Bayesian analysis when the prior information is vague.
- These priors typically produce posteriors with good frequentist properties.
- They are typically improper:Jeffreys priors, reference priors, other benchmark priors.

Non-informative priors

- “Noninformative” or “Benchmark” priors are of interest in Bayesian analysis when the prior information is vague.
- These priors typically produce posteriors with good frequentist properties.
- They are typically improper:Jeffreys priors, reference priors, other benchmark priors. It is necessary to check that the posterior distribution is well-defined (proper).

Non-informative priors

- “Noninformative” or “Benchmark” priors are of interest in Bayesian analysis when the prior information is vague.
- These priors typically produce posteriors with good frequentist properties.
- They are typically improper:Jeffreys priors, reference priors, other benchmark priors. It is necessary to check that the posterior distribution is well-defined (proper).
- There is interest on studying the propriety of the posterior distribution under general prior structures that contain priors obtained by formal rules.

Hierarchy I

- Fernández et al. (1997) proposed the following hierarchical structure:

$$\begin{aligned} \mathbf{u} &\sim p(\mathbf{u}), \\ \varepsilon_j|\sigma_\varepsilon &\sim N(0, \sigma_\varepsilon) \end{aligned} \tag{3}$$

where $p(\mathbf{u})$ is proper. They adopt the following prior structure

$$\pi(\beta, \sigma_\varepsilon) \propto \frac{1}{\sigma_\varepsilon^{b+1}}, \tag{4}$$

where $b \geq 0$. This prior is typically justified as a prior inspired by the structure of the Jeffreys, independence Jeffreys, and reference priors (for different choices of b).

Hierarchy I

- Fernández et al. (1997) proposed the following hierarchical structure:

$$\begin{aligned} \mathbf{u} &\sim p(\mathbf{u}), \\ \varepsilon_j|\sigma_\varepsilon &\sim N(0, \sigma_\varepsilon) \end{aligned} \tag{3}$$

where $p(\mathbf{u})$ is proper. They adopt the following prior structure

$$\pi(\boldsymbol{\beta}, \sigma_\varepsilon) \propto \frac{1}{\sigma_\varepsilon^{b+1}}, \tag{4}$$

where $b \geq 0$. This prior is typically justified as a prior inspired by the structure of the Jeffreys, independence Jeffreys, and reference priors (for different choices of b).

Hierarchy I

- Fernández et al. (1997) proposed the following hierarchical structure:

$$\begin{aligned} \mathbf{u} &\sim p(\mathbf{u}), \\ \varepsilon_j|\sigma_\varepsilon &\sim N(0, \sigma_\varepsilon) \end{aligned} \tag{3}$$

where $p(\mathbf{u})$ is proper. They adopt the following prior structure

$$\pi(\boldsymbol{\beta}, \sigma_\varepsilon) \propto \frac{1}{\sigma_\varepsilon^{b+1}}, \tag{4}$$

where $b \geq 0$. This prior is typically justified as a prior inspired by the structure of the Jeffreys, independence Jeffreys, and reference priors (for different choices of b).

Propriety of the posterior

- Fernández et al. (1997) present the following conditions for the propriety of the corresponding posterior.

  Let $(\mathbf{X} : \mathbf{Z})$ be the entire design matrix.

  1. If $\text{rank}(\mathbf{X} : \mathbf{Z}) < n$, then the posterior distribution exists.
  2. If $\text{rank}(\mathbf{X} : \mathbf{Z}) = n$, then the posterior distribution is improper.

Pros and Cons

- Easy to check conditions.
- The prior is location and scale invariant.
- It allows for the use of any random effects distribution with proper priors on the parameters. This is

$$\begin{aligned} \mathbf{u} &\sim F(\cdot;\theta), \\ \theta &\sim p(\theta). \end{aligned}$$

- The random effects can be assumed to be either dependent or independent.
- The errors distribution can only be assumed to be normal.

Pros and Cons

- Easy to check conditions.
- The prior is location and scale invariant.
- It allows for the use of any random effects distribution with proper priors on the parameters. This is

$$\begin{aligned} \mathbf{u} &\sim F(\cdot;\theta), \\ \theta &\sim p(\theta). \end{aligned}$$

- The random effects can be assumed to be either dependent or independent.
- The errors distribution can only be assumed to be normal.

Pros and Cons

- Easy to check conditions.
- The prior is location and scale invariant.
- It allows for the use of any random effects distribution with proper priors on the parameters. This is

$$\begin{aligned} \mathbf{u} &\sim F(\cdot;\theta), \\ \theta &\sim p(\theta). \end{aligned}$$

- The random effects can be assumed to be either dependent or independent.
- The errors distribution can only be assumed to be normal.

Pros and Cons

- Easy to check conditions.
- The prior is location and scale invariant.
- It allows for the use of any random effects distribution with proper priors on the parameters. This is

$$\begin{aligned} \mathbf{u} &\sim F(\cdot;\theta), \\ \theta &\sim p(\theta). \end{aligned}$$

- The random effects can be assumed to be either dependent or independent.
- The errors distribution can only be assumed to be normal.

Pros and Cons

- Easy to check conditions.
- The prior is location and scale invariant.
- It allows for the use of any random effects distribution with proper priors on the parameters. This is

$$\begin{aligned} \mathbf{u} &\sim F(\cdot;\theta), \\ \theta &\sim p(\theta). \end{aligned}$$

- The random effects can be assumed to be either dependent or independent.
- The errors distribution can only be assumed to be normal.

Extension

- We now focus on the study of “Hierarchy I” under more flexible distributional assumptions on the errors ε_{ij}.

Extension

- We now focus on the study of "Hierarchy I" under more flexible distributional assumptions on the errors ε_{ij}.
- Recall that a scale mixture of normal distributions (SMN_1), with mixing parametric distribution $H(\cdot;\delta)$, is defined as

$$g(x|\mu,\sigma,\delta) = \int_0^\infty \frac{\tau^{1/2}}{(2\pi\sigma)^{1/2}} \exp\left[-\frac{\tau(x-\mu)^2}{2\sigma}\right] dH(\tau;\delta).$$

Extension

- We now focus on the study of "Hierarchy I" under more flexible distributional assumptions on the errors ε_{ij}.
- Recall that a scale mixture of normal distributions (SMN_1), with mixing parametric distribution $H(\cdot;\delta)$, is defined as

$$g(x|\mu,\sigma,\delta) = \int_0^\infty \frac{\tau^{1/2}}{(2\pi\sigma)^{1/2}} \exp\left[-\frac{\tau(x-\mu)^2}{2\sigma}\right] dH(\tau;\delta).$$

This family contains the Student-t distribution with δ degrees of freedom, the Normal, Logistic, Laplace, Power Exponential ... For certain choices of the mixing distribution.

- Consider the hierarchical linear mixed model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon},$$

with hierarchical structure

$$\begin{aligned} \mathbf{u} &\sim p(\mathbf{u}), \\ \varepsilon_j|\sigma_\varepsilon, \delta_\varepsilon &\sim \text{SMN}_1(0, \sigma_\varepsilon, \delta_\varepsilon) \end{aligned}$$

where $p(\mathbf{u})$ is proper, and a certain mixing distribution H_ε. We adopt the following prior structure

$$\pi(\boldsymbol{\beta}, \sigma_\varepsilon, \delta_\varepsilon) \propto \frac{p(\delta_\varepsilon)}{\sigma_\varepsilon^{b+1}},$$

where $b \geq 0$, and $p(\delta_\varepsilon)$ is a proper prior.

- Consider the hierarchical linear mixed model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon},$$

with hierarchical structure

$$\begin{aligned} \mathbf{u} &\sim p(\mathbf{u}), \\ \varepsilon_j | \sigma_\varepsilon, \delta_\varepsilon &\sim \text{SMN}_1(0, \sigma_\varepsilon, \delta_\varepsilon) \end{aligned}$$

where $p(\mathbf{u})$ is proper, and a certain mixing distribution H_ε. We adopt the following prior structure

$$\pi(\boldsymbol{\beta}, \sigma_\varepsilon, \delta_\varepsilon) \propto \frac{p(\delta_\varepsilon)}{\sigma_\varepsilon^{b+1}},$$

where $b \geq 0$, and $p(\delta_\varepsilon)$ is a proper prior.

- Consider the hierarchical linear mixed model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon},$$

with hierarchical structure

$$\begin{aligned} \mathbf{u} &\sim p(\mathbf{u}), \\ \varepsilon_j | \sigma_\varepsilon, \delta_\varepsilon &\sim \text{SMN}_1(0, \sigma_\varepsilon, \delta_\varepsilon) \end{aligned}$$

where $p(\mathbf{u})$ is proper, and a certain mixing distribution H_ε. We adopt the following prior structure

$$\pi(\boldsymbol{\beta}, \sigma_\varepsilon, \delta_\varepsilon) \propto \frac{p(\delta_\varepsilon)}{\sigma_\varepsilon^{b+1}},$$

where $b \geq 0$, and $p(\delta_\varepsilon)$ is a proper prior.

For this model we have the following result.

Theorem

Consider the following conditions:

(a) $\text{rank}(X : Z) < n$,

(b) $b \geq 0$,

(c) $\int_{\Delta_\varepsilon} \int_{\mathbb{R}_+} \tau^{-\frac{b}{2}} p(\delta_\varepsilon) dH_\varepsilon(\tau; \delta_\varepsilon) d\delta_\varepsilon < \infty$.

(d) $\mathbf{y}$ *is not an element of the column space of* $(X : Z)$.

Condition (a) is necessary for the propriety of the posterior.
Conditions (a) – (d) are sufficient for the propriety of the posterior.

- Condition (c) holds for $b = 0$ and any mixing distribution.

- Condition (c) holds for $b = 0$ and any mixing distribution.
- For $b > 0$, a case-by-case analysis is necessary.

- Condition (c) holds for $b = 0$ and any mixing distribution.
- For $b > 0$, a case-by-case analysis is necessary.
- Condition (d) is satisfied with probability 1 since the distributions are continuous.

- Condition (c) holds for $b = 0$ and any mixing distribution.
- For $b > 0$, a case-by-case analysis is necessary.
- Condition (d) is satisfied with probability 1 since the distributions are continuous.
- Propriety for LMM and Stochastic Frontier Models ✓.

- Condition (c) holds for $b = 0$ and any mixing distribution.
- For $b > 0$, a case-by-case analysis is necessary.
- Condition (d) is satisfied with probability 1 since the distributions are continuous.
- Propriety for LMM and Stochastic Frontier Models ✓.
- What about MEAFT?

- Condition (c) holds for $b = 0$ and any mixing distribution.
- For $b > 0$, a case-by-case analysis is necessary.
- Condition (d) is satisfied with probability 1 since the distributions are continuous.
- Propriety for LMM and Stochastic Frontier Models ✓.
- What about MEAFT? Censored observations $\times$.

Theorem

Let $I_1, \ldots, I_{n_c}$ be finite-length intervals on the positive real line, $n_c \leq n$.

Theorem

Let $I_1, \ldots, I_{n_c}$ be finite-length intervals on the positive real line, $n_c \leq n$. Suppose that n_c survival times T_j, $j = 1, \ldots n_c$, are observed as intervals I_j, and the rest of the observations exhibit another type of censoring. Let $(\mathbf{X} : \mathbf{Z})_{n_c}$ represent the design matrix associated to the n_c interval–censored observations.

Theorem

Let $I_1, \ldots, I_{n_c}$ be finite-length intervals on the positive real line, $n_c \leq n$. Suppose that n_c survival times T_j, $j = 1, \ldots n_c$, are observed as intervals I_j, and the rest of the observations exhibit another type of censoring. Let $(\mathbf{X} : \mathbf{Z})_{n_c}$ represent the design matrix associated to the n_c interval–censored observations. Consider the following condition:

(d') *The set $\mathcal{E} = I_1 \times \cdots \times I_{n_c}$ and the column space of $(\mathbf{X} : \mathbf{Z})_{n_c}$ are disjoint.*

Conditions (a)–(c) from the previous Theorem together with (d') are sufficient for the propriety of the posterior.

- MEAFT ✓.

- MEAFT ✓. How do I check (d') in practice?

- MEAFT ✓. How do I check (d') in practice?
- It is possible to show that (d') is equivalent to verifying the infeasibility of the Linear Programming (LP) problem:

$$\begin{aligned} \text{Find} \quad & \max_{\boldsymbol{\eta},\boldsymbol{\xi}} 1, \\ \text{Subject to} \quad & (X : Z)_{n_c}\boldsymbol{\eta} = \boldsymbol{\xi}, \\ \text{and} \quad & \log(l_j) \leq \xi_j \leq \log(u_j), \;\; j = 1, \ldots, n_c. \end{aligned} \tag{5}$$

$\boldsymbol{\eta} \in \mathbb{R}^{p+q}$, $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{n_c}) \in \mathcal{E}$, and $I_j = [l_j, u_j]$, $j = 1, \ldots, n_c$.

- MEAFT ✓. How do I check (d') in practice?
- It is possible to show that (d') is equivalent to verifying the infeasibility of the Linear Programming (LP) problem:

$$\begin{aligned} \text{Find} \quad & \max_{\boldsymbol{\eta},\boldsymbol{\xi}} 1, \\ \text{Subject to} \quad & (X : Z)_{n_c}\boldsymbol{\eta} = \boldsymbol{\xi}, \\ \text{and} \quad & \log(l_j) \leq \xi_j \leq \log(u_j), \;\; j = 1, \ldots, n_c. \end{aligned} \tag{5}$$

$\boldsymbol{\eta} \in \mathbb{R}^{p+q}$, $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{n_c}) \in \mathcal{E}$, and $I_j = [l_j, u_j]$, $j = 1, \ldots, n_c$.
There are several LP solvers available (R, Matlab).

1. We have presented extensions to the standard (with normal assumptions) Bayesian hierarchical linear mixed models with improper priors.

1. We have presented extensions to the standard (with normal assumptions) Bayesian hierarchical linear mixed models with improper priors.
2. These extensions can be used to capture departures from the assumptions of normality in terms of the tail behaviour.

1. We have presented extensions to the standard (with normal assumptions) Bayesian hierarchical linear mixed models with improper priors.
2. These extensions can be used to capture departures from the assumptions of normality in terms of the tail behaviour.
3. Applications.

1. We have presented extensions to the standard (with normal assumptions) Bayesian hierarchical linear mixed models with improper priors.
2. These extensions can be used to capture departures from the assumptions of normality in terms of the tail behaviour.
3. Applications.
4. IMPLEMENTATION.

1. **Thank you for your attention.**

1. **Thank you for your attention.**

1. **Thank you for your attention.**
2. Improper priors on the parameter of the distribution of the random effects?

1. **Thank you for your attention.**
2. Improper priors on the parameter of the distribution of the random effects?The conditions for the propriety of the posterior are restrictive (Hobert and Casella, 1996; Sun et al., 2001; Rubio, 2015).