

Appl. Statist. (2020)
69, Part 5, pp. 1015–1065

Linear mixed effects models for non-Gaussian continuous repeated measurement data

Özgür Asar,

Acıbadem Mehmet Ali Aydınlar University, İstanbul, Turkey

David Bolin,

*King Abdullah University of Science and Technology, Thuwal, Saudi Arabia,
and University of Gothenburg, Sweden*

Peter J. Diggle

Lancaster University, and Health Data Research UK, London, UK

and Jonas Wallin

Lund University, Sweden

[Read before The Royal Statistical Society at a meeting organized by the Research Section on Wednesday, May 13th, 2020, Professor G. P. Nason in the Chair]

Summary. We consider the analysis of continuous repeated measurement outcomes that are collected longitudinally. A standard framework for analysing data of this kind is a linear Gaussian mixed effects model within which the outcome variable can be decomposed into fixed effects, time invariant and time-varying random effects, and measurement noise. We develop methodology that, for the first time, allows any combination of these stochastic components to be non-Gaussian, using multivariate normal variance–mean mixtures. To meet the computational challenges that are presented by large data sets, i.e. in the current context, data sets with many subjects and/or many repeated measurements per subject, we propose a novel implementation of maximum likelihood estimation using a computationally efficient subsampling-based stochastic gradient algorithm. We obtain standard error estimates by inverting the observed Fisher information matrix and obtain the predictive distributions for the random effects in both filtering (conditioning on past and current data) and smoothing (conditioning on all data) contexts. To implement these procedures, we introduce an R package: *ngme*. We reanalyse two data sets, from cystic fibrosis and nephrology research, that were previously analysed by using Gaussian linear mixed effects models.

Keywords: Heavy-tailedness; Latent effects; Longitudinal data; Multivariate analysis; Non-normal distributions; Skewness; Stochastic approximation

1. Introduction

This paper is concerned with the analysis of real-valued *repeated measurement* data that are collected through time: also known as *longitudinal* data. The basic data structure is that repeated measurements of an *outcome variable* are made on each of a number of *subjects* at each of a number of *follow-up times*, which are not necessarily the same for all subjects, with explanatory

Address for correspondence: Jonas Wallin, Lund Universitet Ekonomihogskolan—Statistika, Tycho Brahes vag 1, Lund 220 07, Sweden.
E-mail: jonas.wallin@stat.lu.se

© 2020 The Authors. Journal of the Royal Statistical Society: Series C (Applied Statistics) 0035–9254/20/691015 published by John Wiley & Sons Ltd on behalf of the Royal Statistical Society. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

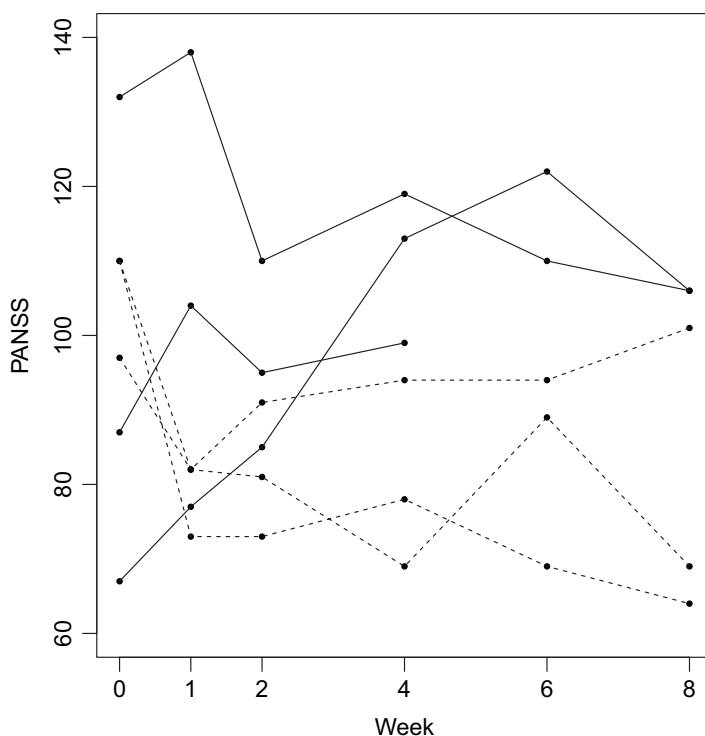


Fig. 1. Data on six patients in a randomized trial of drug therapies for schizophrenia: the outcome variable, PANSS (positive and negative syndrome scale), is a questionnaire-based instrument for assessing the severity of a patient's condition (—, patients from the placebo treatment arm; ----, patients from the active treatment arm)

variables or *covariates* of two kinds also available: *baseline* covariates attached to subjects, and *longitudinal* covariates attached to individual outcomes. We write Y_{ij} for the j th measurement of the outcome on the i th subject, t_{ij} for the corresponding follow-up time, \mathbf{a}_i for the matrix of baseline covariates that are associated with the i th subject and \mathbf{x}_{ij} for the matrix of longitudinal covariates that are attached to the j th measurement on the i th subject.

Fig. 1 shows a simple example, taken from a randomized trial of drug treatments for schizophrenia, in which the outcome variable is a measure of each subject's mental state at times 0, 1, 2, 4, 6 and 8 weeks after randomization to one of two different drug therapies: placebo *versus* active treatment. Here, a_i is a scalar treatment indicator, whereas the general pattern of decreasing responses over time suggests a quadratic trend; hence \mathbf{x}_{ij} consists of t_{ij} and t_{ij}^2 . Fig. 1 shows data from three subjects in each of the two treatment arms; the complete trial included 88 subjects in the placebo group and 435 subjects distributed across five active treatment arms (Henderson *et al.*, 2000). This example shows several features that are typical of studies of this kind: the outcome variable, the PANSS-score (positive and negative syndrome scale) (Kay *et al.*, 1987), is an imperfect measurement instrument for the underlying process of interest, namely each subject's state of mental health at the time of measurement; the outcome variable exhibits stochastic variation both between subjects and between follow-up times within subjects; questions of interest include *estimation* of parameters that define the mean response profiles of the underlying process over time and *prediction* of the trajectory of the process for an individual subject.

Most of the very extensive literature on statistical methods for data of this kind uses either a Gaussian model or, if the inferential goal is restricted to parameter estimation, a set of estimating equations; textbook accounts include Verbeke and Molenberghs (2001), Diggle *et al.* (2002) and Fitzmaurice *et al.* (2011). In this paper, we present methodology for handling repeated measurement data that exhibit long-tailed or skewed departure from Gaussian distributional assumptions.

In Section 2, we review the literature on existing approaches to Gaussian and non-Gaussian modelling of real-valued repeated measurement data. In Section 3, we set out our proposed class of non-Gaussian models. In Section 4, we describe a computationally fast method for likelihood-based inference. Section 5 describes a method for validating the distributional assumptions of the models considered. Section 6 describes two applications. In the first of these, the scientific focus is on estimation of mean response profiles, whereas in the second the focus is on realtime individual level prediction. Section 7 presents the results from two simulation studies and Section 8 describes our R package, `ngme`, that implements the new methodology. In Section 9, we discuss some potential extensions, including models for categorical or count data (Molenberghs and Verbeke, 2005) and joint modelling of repeated measurement and time-to-event data (Rizopoulos, 2012). Technical details are presented in the appendices.

2. Literature review

2.1. Gaussian models for real-valued repeated measurement data

Laird and Ware (1982) were the first to consider modelling repeated measurements as noisy versions of underlying signals that can be decomposed into fixed effects, $\mathbf{a}_i^T \boldsymbol{\alpha} + \mathbf{x}_{ij}^T \boldsymbol{\beta}$, and random effects, $\mathbf{d}_{ij}^T \mathbf{U}_i$, leading to the mixed effects model

$$Y_{ij} = \mathbf{a}_i^T \boldsymbol{\alpha} + \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{d}_{ij}^T \mathbf{U}_i + \sigma Z_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, m, \quad (1)$$

where n_i is the number of measurements on the i th subject, m is the number of subjects, the individual level \mathbf{U}_i are mutually independent, zero-mean multivariate normal, $\mathbf{U}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, and the Z_{ij} are mutually independent $N(0, 1)$.

A widely used special case of model (1) is the ‘random-intercept and random-slope’ model in which each subject’s random effect is a linear function of time. This model is very useful when the data contain only a small number of repeated measurements per individual. With longer sequences, the assumption that individual random-effect trajectories can be approximated by straight lines becomes implausible, because of non-linearities in the trajectories. Diggle (1988) proposed adding to the model a *time-varying* random-effect term $W_i(t)$, specified as a stationary stochastic process. Taylor *et al.* (1994) and Diggle *et al.* (2015) later considered non-stationary options for $W_i(t)$. The general specification for models of this kind is that

$$Y_{ij} = \mathbf{a}_i^T \boldsymbol{\alpha} + \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{d}_{ij}^T \mathbf{U}_i + W_i(t_{ij}) + \sigma Z_{ij}, \quad (2)$$

where, in addition to the notation that has already been introduced, the $W_i(t)$ are independent copies of a continuous time zero-mean Gaussian process with covariance function $\gamma(t, t') = \text{cov}\{W_i(t), W_i(t')\}$. We consider the elements of both the \mathbf{a}_i and \mathbf{x}_{ij} to be prespecified constants. This implicitly assumes, in particular, that, if any time-varying covariate is not prespecified, it is stochastically independent of all other terms in the model; hence conditioning on it is innocuous. We can then drop the term $\mathbf{a}_i^T \boldsymbol{\alpha}$ in model (2) by allowing elements of \mathbf{x}_{ij} to take identical values for all j that are associated with any fixed i . For the covariance function $\gamma(t, t')$, we use the stationary Matérn (1960) family:

$$\gamma(t, t') = \omega^2 \{2^{\phi-1} \Gamma(\phi)\}^{-1} (\kappa |t - t'|)^{\phi} K_{\phi}(\kappa |t - t'|), \quad (3)$$

where $\Gamma(\phi)$ is the complete gamma function, $\omega^2 > 0$ denotes the variance, $\phi > 0$ is a shape parameter, $\kappa > 0$ is a scale parameter and K_{ϕ} is a modified Bessel function of the second kind of order ϕ . The corresponding Gaussian process $W_i(t)$ is $\lceil \phi \rceil - 1$ times mean-square differentiable, where ' $\lceil \cdot \rceil$ ' denotes the ceiling function. An alternative way of capturing non-linear behaviour of repeated measurements is to specify the random effects as regression splines or polynomials with stochastic coefficients (Fitzmaurice *et al.* (2011), chapter 19). We do not consider these approaches in this paper, since they appear to us less natural than the stochastic process approach and would require many more parameters to achieve the same flexibility in shape. That said, there are connections between an integrated random-walk process and a smoothing spline representation of the $W_i(t)$ (Wahba, 1990; Zhu and Dunson, 2017).

Likelihood-based inference for model (2) is straightforward. The likelihood function is a product of m multivariate normal densities with dimensions n_i . For typical study designs, the n_i are sufficiently small that the required matrix calculations are not computationally demanding.

In the continuous time setting, it is helpful to exploit an alternative representation of a Gaussian process $W(\cdot)$ as the solution to a stochastic differential equation,

$$\mathcal{D}W(t) = dL(t), \quad (4)$$

where \mathcal{D} is a differential operator and $dL(t)$ is continuous time Gaussian white noise (Lindgren *et al.*, 2011). For example, the integrated random-walk model used by Diggle *et al.* (2015) and Zhu and Dunson (2017) corresponds to $\mathcal{D} = \partial^2 / \partial t^2$, whereas the Matérn model corresponds to

$$\mathcal{D} = \left(\kappa^2 - \frac{\partial^2}{\partial t^2} \right)^{(2\phi+1)/4}. \quad (5)$$

For the stochastic differential equation representation of an integrated Ornstein–Uhlenbeck process, see Zhu *et al.* (2011a, b).

In applications for which only the regression parameters β are of scientific interest, estimating equations offer an alternative to likelihood-based estimation. In the current context, this approach was introduced by Liang and Zeger (1986), working in the wider setting of generalized linear models. For linear models, the approach consists of estimating β by weighted least squares; hence

$$\tilde{\beta} = \left(\sum_{i=1}^m \mathbf{x}_i^T \mathbf{F}_i \mathbf{x}_i \right)^{-1} \sum_{i=1}^m \mathbf{x}_i^T \mathbf{F}_i \mathbf{Y}_i, \quad (6)$$

where, for each i , $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})^T$, \mathbf{x}_i is the $n_i \times k$ matrix whose j th row is \mathbf{x}_{ij}^T and the \mathbf{F}_i are weight matrices. Rewriting equation (6) in an obvious shorthand notation as $\tilde{\beta} = \mathbf{D}\mathbf{Y}$, inference for β uses the result that $\tilde{\beta}$ is asymptotically multivariate Gaussian with mean β and variance $\mathbf{D}\mathbf{C}\mathbf{D}^T$, where $\mathbf{C} = \text{var}(\mathbf{Y})$, a block diagonal matrix with non-zero blocks, $\mathbf{C}_i = \text{var}(\mathbf{Y}_i)$. If $\mathbf{F}_i = \mathbf{C}_i^{-1}$, then $\tilde{\beta}$ is the maximum likelihood estimator for β .

The basic idea behind equation (6) is to choose, rather than to estimate, a set of matrices \mathbf{F}_i that reflect a reasonable *working covariance structure* for the matrices $\mathbf{C}_i = \text{var}(\mathbf{Y}_i)$, but not to rely on the correctness of the chosen structure. Instead, the unknown matrix \mathbf{C}_i is replaced by a non-parametric estimate $\tilde{\mathbf{C}}_i$. One such set of estimates is given by $\tilde{\mathbf{C}}_i = n_i^{-1} (\mathbf{Y}_i - \mathbf{x}_i \tilde{\beta})(\mathbf{Y}_i - \mathbf{x}_i \tilde{\beta})^T$. Individually, each $\tilde{\mathbf{C}}_i$ is a very poor estimate of \mathbf{C}_i , but the implicit averaging in equation (6) leads to consistent estimation of $\text{var}(\tilde{\beta})$ in the limit $m \rightarrow \infty$ for fixed n_i (Liang and Zeger, 1986).

2.2. Non-Gaussian models for real-valued repeated measurement data

The existing literature on non-Gaussian models takes as its starting point a linear model with correlated errors:

$$Y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + S_{ij}, \quad (7)$$

where, in the case of a common set of follow-up times, t_1, \dots, t_n , for each subject, the $\mathbf{S}_i = (S_{i1}, \dots, S_{in})^T$ are independent copies of a zero-mean multivariate normal random vector (Jennrich and Schluchter, 1986). Most references consider the Laird–Ware approach as presented in model (1), where

$$S_{ij} = \mathbf{d}_{ij}^T \mathbf{U}_i + Z_{ij}. \quad (8)$$

Liu and Rubin (1995), Lange *et al.* (1989) and Pinheiro *et al.* (2001) replaced each S_{ij} in equation (7) or (8) by $S_{ij}^* = S_{ij}/\sqrt{V_i}$, where the V_i are mutually independent unit mean gamma-distributed random variables. They estimated the model parameters by maximum likelihood using an expectation–maximization algorithm (Dempster *et al.*, 1977). Lin and Wang (2011) considered Bayesian methods of inference for the same class of models. Matos *et al.* (2013) extended the work of Pinheiro *et al.* (2001) to allow censored outcomes.

Song *et al.* (2007) and Zhang *et al.* (2009) considered an extension to Lange *et al.* (1989) by allowing the gamma-distributed scaling factor V_i to apply to either one of the two components on the right-hand side of equation (8). Lin and Lee (2007) applied the gamma-distributed scaling factor only to the random-effects term, $\mathbf{d}_{ij}^T \mathbf{U}_i$, but also replaced the mutually independent Z_{ij} by a set of auto-regressive processes; this restricts its applicability to data with equally spaced measurement times.

Rosa *et al.* (2003) and Tian *et al.* (2008) also used the formulation $S_{ij}^* = S_{ij}/\sqrt{V_i}$, but without restricting the V_i to be gamma distributed. Lange and Sinsheimer (1993) called the resulting family of distributions the *normal–independent* family, a special case of which is a mixture of normal distributions. The R package *heavy* (Osorio, 2016) fits this class of models. In a series of papers, V. H. Lachos and colleagues have developed methodology for fitting non-linear mixed models by using the normal–independent family; see Lachos *et al.* (2009, 2010, 2011, 2012, 2013) Zeller *et al.* (2010) and Cabral *et al.* (2012) and also independent contributions by Verbeke and Lesaffre (1996), Sun *et al.* (2008), Ho and Lin (2010), De la Cruz (2014), Zhang *et al.* (2015) and Yavuz and Arslan (2016).

Several researchers have extended the single-term modelling framework (8) by decoupling the scalings of the random effects and the measurement error terms. See, for example, Rosa *et al.* (2004), Aralleno-Valle *et al.* (2007), Jara *et al.* (2008), Meza *et al.* (2012), Choudhary *et al.* (2014) and Bai *et al.* (2016). Lu and Zhang (2014) extended the approach to include non-ignorable drop-out.

Wang and Fan (2011, 2012), Lin and Wang (2013) and Kazemi *et al.* (2013) used the normal–independent family to model multivariate repeated measurement data.

Others have taken a semiparametric approach to the problem, for example by using a Dirichlet process prior for the random effects or leaving the random-effects distribution unspecified. See Kleinman and Ibrahim (1998), Ghidry *et al.* (2004), Tao *et al.* (2004), Subtil and Rabilloud (2010), Davidian and Gallant (1993), Zhang and Davidian (2001) and Vock *et al.* (2012). Koller (2016) considered robust estimating equations.

We have found only two references that considered the general form of model (2) with three stochastic components with the single-term formulation (8), namely Stirrup *et al.* (2015) and Asar *et al.* (2016), and none that allows the three scaling factors to be decoupled.

3. A flexible class of non-Gaussian models

Our aim in this section is to set out a version of the mixed effects model

$$Y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{d}_{ij}^T \mathbf{U}_i + W_i(t_{ij}) + \sigma Z_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, m, \quad (9)$$

that, we believe for the first time, allows Gaussian or non-Gaussian distributional specifications of the three stochastic components \mathbf{U}_i , $W_i(t)$ and Z_{ij} to be decoupled.

Writing \mathbf{B} and \mathbf{H} to denote generic vector-valued random variables, we replace the Gaussian assumption for each of the components with a normal variance–mean mixture of the form

$$\mathbf{B} = \boldsymbol{\delta} + \boldsymbol{\mu}V + \sqrt{V}\boldsymbol{\Sigma}^{1/2}\mathbf{H}, \quad (10)$$

where $\boldsymbol{\delta}$ and $\boldsymbol{\mu}$ are parameter matrices, $\mathbf{H} \sim N(\mathbf{0}, \mathbf{I})$ with \mathbf{I} being the identity, and V is a random variable that takes values on \mathbb{R}^+ . We need to impose some restrictions on the distribution of V for the inferential algorithms that we develop in Section 4 to be practicable. For the subject-specific random effect, \mathbf{U}_i , and the measurement-specific noise, Z_{ij} , the only necessary restriction is that V has a known distribution. However, to simplify parameter estimation, we shall impose the additional restriction that $V|\mathbf{H}$ also has a known distribution. For the subject-specific continuous time stochastic process, $W_i(t)$, we use a numerical discretization of the differential operator (4) to generate realizations of the process. For this reason, we need the distribution to be closed under arbitrary discretization, which we ensure by requiring that the distribution of V be closed under convolution. Our specific proposals for \mathbf{U}_i , $W_i(t)$ and Z_{ij} are described below in more detail.

A flexible choice for \mathbf{B} is the multivariate generalized hyperbolic (GH) distribution (Barndorff-Nielsen, 1977; Vilca *et al.*, 2014). This distribution can be generated from the mixture representation (10) by specifying a generalized inverse Gaussian (GIG) distribution for V . The density function of the GIG distribution is

$$f(x; p, a, b) = \frac{(a/b)^{p/2}}{2K_p(\sqrt{ab})} x^{p-1} \exp\left(-\frac{a}{2}x - \frac{b}{2}x^{-1}\right), \quad (11)$$

where K_p is the modified Bessel function of the third kind, of order p , whereas a and b are positive-valued parameters. We denote this distribution by $\text{GIG}(p, a, b)$ and refer the reader to Jørgensen (1982) for more details. An important property of this distribution is that, for any $c > 0$, $cV \sim \text{GIG}(p, a/c, cb)$. Another property that is useful for the construction of the sampling-based inferential algorithms that we introduce in Sections 4.2 and 4.4 is that the conditional distribution of V given the observed data is also GIG.

The GH distribution includes several widely used distributions as special cases, e.g. the Student t , generalized asymmetric Laplace (GAL), normal–inverse Gaussian (NIG) and Cauchy distributions. Specific parameter configurations for the distributions of V that give each of these special cases are presented in Table 1. Note that, for both the NIG and the GAL distributions, the formulation is overparameterized for \mathbf{B} in equation (10). One therefore needs to fix a or b .

3.1. Noise and random effects

Since the measurement noise is univariate, we can write the mixture representation (10) as

$$Z_{ij} = \delta^Z + \mu^Z V_{ij}^Z + \sigma \sqrt{V_{ij}^Z} Z_{ij}^*, \quad (12)$$

where $Z_{ij}^* \sim N(0, 1)$. To maintain the interpretation of σ^2 as the variance of the noise, at least in the symmetric case, we constrain the values of the GIG parameters a , b and p , so that $E[V_{ij}^Z] = 1$,

Table 1. Some special cases of the GH distribution, their mixing distributions and their corresponding GIG forms†

Distribution of \mathbf{B}	Mixing distribution of V	GIG form of the mixing distribution
t	$\text{IGam}(\nu/2, \nu/2)$	$\text{GIG}(\nu/2, \nu, 0)$
NIG	$\text{IG}(a, b)$	$\text{GIG}(-\frac{1}{2}, a, b)$
GAL	$\text{Gam}(p, a)$	$\text{GIG}(p, 2a, 0)$
Cauchy	$\text{IGam}(\frac{1}{2}, b/2)$	$\text{GIG}(-\frac{1}{2}, 0, b)$

†Gam indicates the gamma family of distributions, IGam the inverse gamma family and IG the inverse Gaussian family.

and hence $\text{var}(\sigma\sqrt{V_{ij}^Z}Z_{ij}^*) = \sigma^2$. We further set $\delta^Z = -\mu^Z$ to ensure that the measurement noise is zero mean, i.e. $E[Z_{ij}] = 0$.

An alternative to representation (12) is to attach a single random variable V_i^Z to all of the noise terms Z_{ij} on the i th subject, i.e. $Z_{ij} = \delta^Z + \mu^Z V_i^Z + \sqrt{V_i^Z} \sigma Z_{ij}^*$. The distribution of V_i can then be interpreted as a random effect for patient-specific measurement noise variance. Note, in particular, that this introduces stochastic dependence between Z_{ij} and $Z_{ij'}$ for $j \neq j'$, although they are conditionally independent given V_i^Z .

For the random effects, we let $\mathbf{U}_i = \delta^U + \mu^U V_i^U + \sqrt{V_i^U} \Sigma^{1/2} \mathbf{U}_i^*$, where V_i is a unit mean GIG random variable and $\mathbf{U}_i^* \sim N(\mathbf{0}, \mathbf{I})$ with \mathbf{I} as before. We again set $\delta^U = -\mu^U$ to ensure that $E[\mathbf{U}_i] = \mathbf{0}$.

3.2. Stochastic process

A simple way to introduce a non-Gaussian stochastic process term in expression (9) would be again to include a subject-specific scaling, i.e. $W_i(t) = V_i^W W_i^*(t)$, where V_i^W follows a unit mean GIG distribution. However, this approach would not be able to capture interesting within-subject departures from Gaussian behaviour, e.g. jumps or asymmetries in the sample paths of $W_i(t)$. To provide the required flexibility, we instead use non-Gaussian generalizations of the stochastic differential equation (4). Specifically, we propose modelling the $W_i(t)$ as independent copies of the solution to

$$\mathcal{D}W_i(t) = dL_i(t), \quad (13)$$

where the L_i are independent copies of a non-Gaussian Lévy process, i.e. a process with independent and stationary increments. In practice, we work with a discretized version of equation (13), for which Bolin (2014) showed that a type G Lévy process for $L_i(t)$ is a suitable candidate. The implication is that the increments of L_i have a distribution that corresponds to the specification given by equation (10).

One approach would therefore be to choose the distribution of V^W as a GIG distribution, which would yield the GH processes of Eberlein (2001). However, as noted earlier, we require the distribution of V^W to be closed under convolution (Wallin and Bolin, 2015). Also, the stochastic gradient method for parameter estimation that we introduce in Section 4 requires the ability to sample from the conditional distribution of V^W given all other components in the model. Within the GH family, the NIG, GAL and Cauchy distributions are the only ones that meet these requirements (Podgórski and Wallin, 2016). In Table 2, we present the parameterization of the mixing distribution for the three cases. Using any of these distributions for the increments of L_i

Table 2. The parameterizations of the three types of GH processes satisfying closure under convolution

<i>Distribution of $L(t)$</i>	<i>Mixing distribution of V^W</i>	<i>GIG form of the mixing distribution</i>
Cauchy	$\text{IGam}(\frac{1}{2}, t^2/2)$	$\text{GIG}(-\frac{1}{2}, 0, t^2)$
NIG	$\text{IG}(a, bt^2)$	$\text{GIG}(-\frac{1}{2}, a, bt^2)$
GAL	$\text{Gam}(pt, a/2)$	$\text{GIG}(pt, a, 0)$

in equation (13) results in models with the same covariance structure as if L_i were Gaussian, but with more general marginal distributions. The NIG choice makes L_i an NIG process (Barndorff-Nielsen, 1997a). This class of processes has been used in financial modelling; see Barndorff-Nielsen (1997b), Bibby and Sørensen (2003), Tankov (2003) and Eberlein (2001), for more details.

3.2.1. The choice of operator

As previously mentioned, using \mathcal{D} as in equation (5) yields a process with a Matérn covariance function. More specifically, if the process is defined on the entire real line, then it has Matérn covariance. In practice, we restrict the process to a bounded temporal interval and impose boundary conditions on the operator to obtain a well-posed problem. Common choices for these artificial boundary conditions are either homogeneous Neumann or homogeneous Dirichlet conditions. The effect of these artificial boundary conditions is small for distances that are larger than twice the practical correlation range of the process (Lindgren *et al.*, 2011). We therefore define the process on an extended temporal domain $\tilde{T} = [-r, t_{\max} + r]$ where all measurement times lie within the interval $[0, t_{\max}]$ and r is a value that is larger than the practical correlation range.

For $\phi = \frac{1}{2}$, the operator (5) is $\mathcal{D}_1 = (\kappa^2 - \partial^2/\partial t^2)^{1/2}$, which results in a process with an exponential covariance function $E[W(t)W(t+h)] = (2\kappa)^{-1} \exp(-\kappa|h|)$. Another, perhaps more natural, choice that results in a process with exponential covariance when defined on the entire real line is $\mathcal{D}_2 = \kappa + \partial/\partial t$. This can easily be seen by computing the power spectrum, $S_W(\omega)$, of the corresponding stochastic process. Taking the Fourier transform of the stochastic differential equation gives $(\kappa + i\omega)\hat{W}(\omega) = d\tilde{L}(\omega)$. Using the fact that the power spectrum of $d\tilde{L}(\omega)$, is constant, it follows that

$$S_W(\omega) \propto \{(\kappa + i\omega)(\kappa - i\omega)\}^{-1} = (\kappa^2 + \omega^2)^{-1},$$

which we recognize as the spectral density corresponding to the exponential covariance function. This result is well known, since in the Gaussian case the process is the classical Ornstein–Uhlenbeck process. When posed on the bounded domain \tilde{T} , we only have to equip \mathcal{D}_2 with one boundary condition. A natural choice in this case is a stochastic Dirichlet boundary condition at either end point, $W(-r) = W$ or $W(t_{\max} + r) = W$, where W is a random variable with distribution equal to the marginal distribution for $W(t)$, when the model is formulated on the entire real line. This results in a stationary model and there is no need to extend the domain of interest.

When the driving noise of the process is Gaussian, the models that are formulated by using \mathcal{D}_1 and \mathcal{D}_2 are equivalent in distribution (apart from the behaviour at the boundary of the domain). However, for non-Gaussian models the processes that are formulated by using the

two choices are *not* equivalent: the kernel of \mathcal{D}_1 is symmetric, whereas the kernel of \mathcal{D}_2 is completely asymmetric. This affects the appearance of trajectories of the process. For \mathcal{D}_1 , the trajectories are symmetric in time, i.e. the process is time reversible, whereas the trajectories can be asymmetric when \mathcal{D}_2 is used. The difference between the symmetric and asymmetric models is illustrated in Fig. 2, where the same driving NIG noise is used to simulate a trajectory using \mathcal{D}_1 and \mathcal{D}_2 . Using \mathcal{D}_2 allows for sharp jumps in the trajectories that are not present when \mathcal{D}_1 is used.

One way to construct a model with a general Matérn covariance that allows for asymmetries in the sample paths is to add a fractional exponent to \mathcal{D}_2 , to give $\mathcal{D}_3 = (\kappa + \partial/\partial t)^{(2\phi+1)/2}$. This results in a process with a Matérn covariance, but with asymmetry in the trajectories depending on where boundary conditions are imposed. It should be noted that these models have Markov properties when $\phi = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \dots$, which simplifies simulation as explained in the next section.

Besides the Matérn models, another option that is used for longitudinal data is the integrated random-walk model (Diggle *et al.*, 2015). This can be seen as a special case of \mathcal{D}_3 with $\kappa = 0$ and $\phi = \frac{3}{2}$, and can thus be handled in the same way as the Matérn models.

3.2.2. Discretization

We need to discretize time to use the stochastic differential equation (13). For this, we use the approximation

$$W(t) = \sum_{k=1}^K \phi_k(t) W_k, \quad (14)$$

where $\mathbf{W} = (W_1, \dots, W_K)^T$ is a vector of random variables and the $\phi_k(t)$ are basis functions. We use a set of piecewise linear basis functions such that

$$\phi_1(t) = \begin{cases} 1 - \frac{t-s_1}{s_2-s_1}, & s_1 < t < s_2, \\ 0, & \text{otherwise,} \end{cases}$$

$$\phi_K(t) = \begin{cases} \frac{t-s_{K-1}}{s_K-s_{K-1}}, & s_{K-1} < t < s_K, \\ 0, & \text{otherwise,} \end{cases}$$

and, for $k = 2, 3, \dots, (K-1)$,

$$\phi_k(t) = \begin{cases} \frac{t-s_{k-1}}{s_k-s_{k-1}}, & s_{k-1} < t < s_k, \\ 1 - \frac{t-s_k}{s_{k+1}-s_k}, & s_k < t < s_{k+1}, \\ 0, & \text{otherwise,} \end{cases}$$

where $-r = s_1 < s_2 < \dots < s_{K-1} < s_K = t_{\max} + r$. In the case of a Dirichlet boundary condition at $t = -r$, the function ϕ_1 is removed from the expansion, whereas, in the case of the Dirichlet boundary condition at $t = t_{\max} + r$, ϕ_K is removed. The distribution of the stochastic weights is computed by using either a Galerkin finite element discretization or a Petrov–Galerkin method depending on the operator; details are given in Appendix B. The result for the non-Gaussian case can be written as

$$\mathbf{W} | \mathbf{V}^W \sim N[\mathbf{K}^{-1} \{ \mathbf{h}^T \delta^W + (\mathbf{V}^W)^T \boldsymbol{\mu}^W \}, \mathbf{K}^{-1} \text{diag}(\mathbf{V}^W) (\mathbf{K}^{-1})^T], \quad (15)$$

where \mathbf{K} is a matrix corresponding to a discretization of the differential operator and $V_k^W \sim \text{IG}(\nu, h_k^2 \nu)$ where h_k are fixed constants depending on the basis functions. Note that we again set $\delta^W = -\boldsymbol{\mu}^W$ to satisfy $E[W(t)] = 0$.

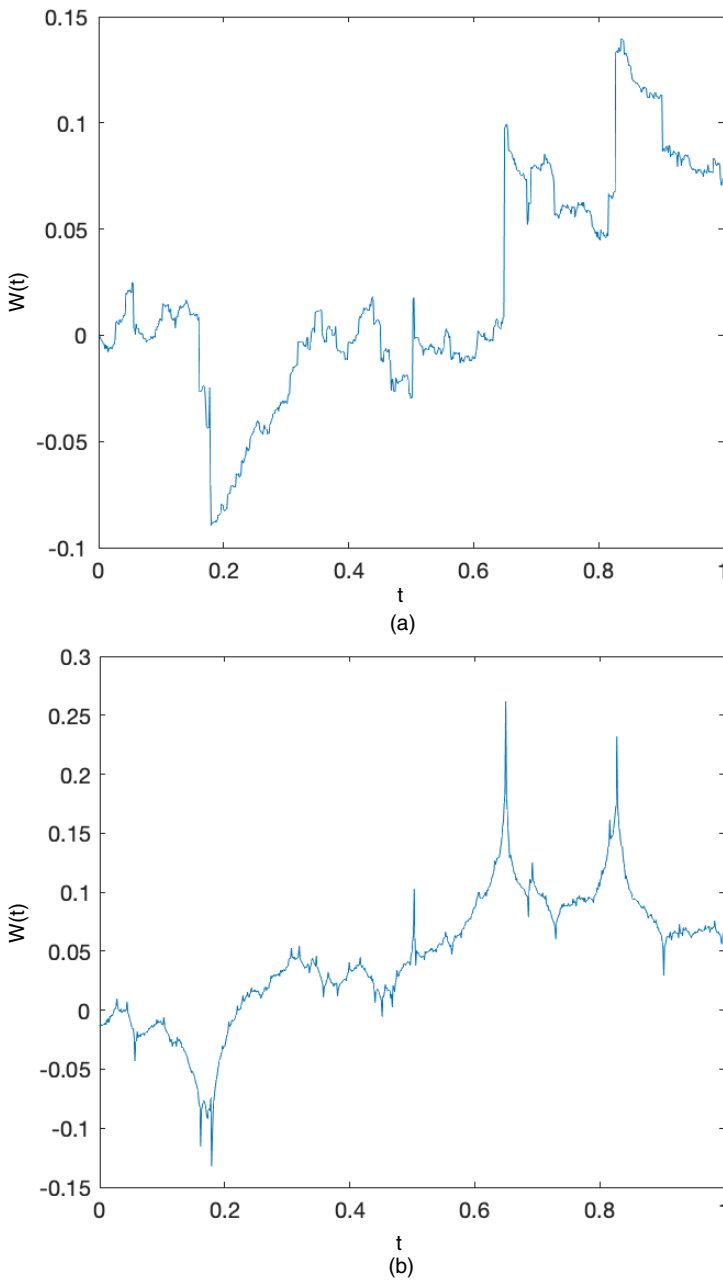


Fig. 2. Simulation of an NIG process with exponential covariance function using the operator (a) \mathcal{D}_2 and (b) \mathcal{D}_1 , with the same driving noise in both cases: one can note that the trajectory is asymmetric in (a)

Note that the parameter ν controls the tails of the marginal distribution of the process. The limiting case $\nu \rightarrow 0$ is the Cauchy process, whereas the limiting case $\nu \rightarrow \infty$ is a Gaussian process. These are exactly the properties that we need to use our likelihood-based methods to assess whether a standard, and undeniably convenient, Gaussian assumption for any or all of the stochastic components of expression (9) is adequate.

4. Likelihood-based inference

4.1. Hierarchical representation

Our specification of a normal variance–mean mixture for each of the stochastic components of expression (9) makes likelihood-based inference practicable via the following hierarchical representation of the model. For subject i , let \mathbf{V}_i^Z , V_i^U and \mathbf{V}_i^W denote the stochastic variance factors corresponding to the noise, random-effects and stochastic process components of expression (9), and write $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{ini})^T$ for the corresponding set of repeated measurements. Let $\mathbf{W}_i = \{W_{ik} : k = 1, \dots, K\}$ be the stochastic weight vector for the i th subject in the approximation of $W_i(t)$ given by equation (14), and \mathbf{A}_i the $n_i \times K$ matrix with (j, k) th element $\phi_k(t_{ij})$. Write \mathbf{x}_i and \mathbf{d}_i for the matrices with j th rows \mathbf{x}_{ij}^T and \mathbf{d}_{ij}^T respectively. Finally, let Θ denote the complete set of model parameters. The model for the i th subject then has the following hierarchical representation:

$$\begin{aligned} \mathbf{Y}_i | \mathbf{W}_i, \mathbf{U}_i, \mathbf{V}_i^Z &\sim N\{\mathbf{x}_i^T \beta + \mathbf{d}_i^T \mathbf{U}_i + \mathbf{A}_i \mathbf{W}_i + (-1 + \mathbf{V}_i^Z) \mu^Z, \sigma^2 \text{diag}(\mathbf{V}_i^Z)\}, \\ \mathbf{U}_i | V_i^U &\sim N(-\mu^U + \mu^U V_i^U, V_i^U \Sigma), \\ \mathbf{W}_i | V_i^W &\sim N\{\mathbf{K}^{-1}(-\mu^W \mathbf{h} + \mu^W \mathbf{V}_i^W), \mathbf{K}^{-1} \text{diag}(\mathbf{V}_i^W)(\mathbf{K}^{-1})^T\}, \end{aligned}$$

coupled with a final layer in the hierarchy: the distributions of the stochastic variance factors, which are GIG with parameters that depend on the model choice. By integrating out the latent variables and variance components, these equations collectively determine the contribution of the i th subject to the marginal log-likelihood $L(\Theta; \mathbf{Y}_i)$. As the vectors \mathbf{Y}_i from the m subjects are independent, the overall log-likelihood is

$$L(\Theta; \mathbf{Y}) = \sum_{i=1}^m L_i(\Theta; \mathbf{Y}_i).$$

4.2. Stochastic gradient estimation

The computations that are required for maximum likelihood estimation are cumbersome for problems that involve longitudinal data sets with large numbers of subjects and repeats, even using the computationally efficient approximation (14). Our proposed algorithm for maximum likelihood estimation therefore uses a stochastic gradient method that calculates the gradient of the objective function at each step of the maximization by subsampling.

A stochastic gradient method for the general problem of minimizing an objective function $f(\Theta)$ starts with an initial guess $\Theta^{(0)}$, and then iteratively updates Θ according to

$$\Theta^{(n+1)} = \Theta^{(n)} - \eta_n Q_n(\Theta^{(n)}), \quad (16)$$

where $Q_n(\Theta)$ is a random variable such that $E[Q_n(\Theta)] = \nabla_{\Theta} f(\Theta)$ and η_n is a sequence of positive numbers such that $\sum_{n=1}^{\infty} \eta_n = \infty$ and $\sum_{n=1}^{\infty} \eta_n^2 < \infty$; an example is $\eta_n \propto 1/\eta^n$ with $0.5 < \eta \leq 1$. Under mild regularity conditions, the resulting sequence $\Theta^{(n)}$ converges to a stationary point of $f(\Theta)$ (Kushner and Yin, 2003; Andrieu *et al.*, 2007).

For maximum likelihood estimation, $f(\Theta) = -L(\Theta; \mathbf{Y})$. If the data set contains a large number of subjects we use only a small, randomly sampled subset in each iteration to generate a computationally efficient stochastic gradient method. For this, $-\nabla_{\Theta} L(\Theta; \mathbf{Y})$ can be replaced by the random variable

$$Q_n(\Theta) = -\nabla_{\Theta} L_s(\Theta; \mathbf{Y}) = -s \sum_{i=1}^m \nabla_{\Theta} L_i(\Theta; \mathbf{Y}_i) J_i, \quad (17)$$

where the J_i are independent Bernoulli random variables with $P(J_i = 1) = 1/s$. Since the expected value of $\nabla_{\Theta} L_s(\Theta; \mathbf{Y})$ with respect to these random variables is equal to $\nabla_{\Theta} L(\Theta; \mathbf{Y})$ for any s , the resulting stochastic gradient method (16) will converge to a stationary point of the log-likelihood. Our experience, for example with the two case-studies that we describe in Section 6, has been that, for data sets containing a large number of subjects, often we need only a small proportion of the available measurement sequences \mathbf{Y}_i at each iteration to estimate the parameters reliably. For example, we used only 688 (3%) subjects out of 22910 for the renal case-study that is presented in Section 6.2.

For our non-Gaussian models, an additional complication is that the likelihood is not available in an explicit form. However, using Fisher's identity (Dempster *et al.*, 1977) we can compute the gradient of the log-likelihood without computing the log-likelihood itself. For all versions of our model, the log-likelihood conditional on the variance components $\mathbf{V} = \{\mathbf{V}_i^W, \mathbf{V}_i^U, \mathbf{V}_i^Z\}_{i=1}^m$ is Gaussian and thus explicit. Fisher's identity then gives

$$\nabla_{\Theta} L_s(\Theta; \mathbf{Y}) = E_{\mathbf{V}}\{\nabla_{\Theta} L_s(\Theta; \mathbf{Y}, \mathbf{V})|\mathbf{Y}\},$$

where $L_s(\Theta; \mathbf{Y}, \mathbf{V})$ is the log-likelihood augmented with \mathbf{V} , which is explicitly available since $\mathbf{Y}|\mathbf{V}$ is Gaussian and \mathbf{V} is GIG. The expectation with respect to \mathbf{V} is not, in general, explicit but can be approximated by Monte Carlo (MC) sampling from the conditional distribution $\mathbf{V}|\mathbf{Y}$. We use a Gibbs sampler and iterate between sampling from the conditional distributions $\mathbf{V}|\mathbf{U}, \mathbf{W}, \mathbf{Y}$ and $\mathbf{U}, \mathbf{W}|\mathbf{V}, \mathbf{Y}$, where $\mathbf{U} = \{\mathbf{U}_i\}_{i=1}^m$ and $\mathbf{W} = \{\mathbf{W}_i\}_{i=1}^m$; for details, see Appendix A. Convergence of algorithms of this kind was studied in Andrieu *et al.* (2007).

When using stochastic gradient optimization to maximize over many parameters, it is important to scale the gradient by a preconditioner to give a Newton-like iteration. Our proposed algorithm therefore is

$$\Theta^{(n+1)} = \Theta^{(n)} - \eta_n \mathbf{I}(\Theta^{(n)})^{-1} Q_n(\Theta^{(n)}), \quad (18)$$

where $\mathbf{I}(\Theta^{(n)})^{-1}$ is a preconditioner to be determined and $Q_n(\Theta^{(n)})$ is a stochastic approximation of the gradient based on subsampling and MC integration over \mathbf{V} using the Gibbs sampler. One option for the preconditioner is $\mathbf{I}^*(\Theta) = -E_{\mathbf{V}}[\nabla_{\Theta}^2 L_s(\Theta; \mathbf{Y}, \mathbf{V})|\mathbf{Y}]$. Calculation of $\mathbf{I}^*(\Theta)$ is typically easy, since $\nabla_{\Theta}^2 L_s(\Theta; \mathbf{Y}, \mathbf{V})$ is often explicit and can be calculated simultaneously with the gradient. Lange (1995) described the connection between using $\mathbf{I}^*(\Theta)$ and the expectation-maximization algorithm. However, if the same variables are used for the MC estimates of the expectations in $\mathbf{I}^*(\Theta)$ and $Q_n(\Theta)$, the joint updating step (18) will be biased because of correlation between the two estimated expectations. One way to avoid this is to use different samples of $\mathbf{V}|\mathbf{Y}$ to compute the two expectations. Instead we use a preconditioner that is similar to the *complete* Fisher information, cFIM,

$$\mathbf{I}_{\text{cFIM}}(\Theta) = -E_{\mathbf{V}, \mathbf{Y}}[\nabla_{\Theta}^2 L_s(\Theta; \mathbf{Y}, \mathbf{V})], \quad (19)$$

which often can be computed explicitly. Note that, in equation (19), the expectation is taken over both \mathbf{Y} and \mathbf{V} . Since the expectation is not conditioned on \mathbf{Y} , cFIM does not suffer from the same biasedness issues as $\mathbf{I}^*(\Theta)$. However, it can still be biased if subsampling is used, and we therefore use a preconditioner that is a weighted average of $\mathbf{I}_{\text{cFIM}}(\Theta)$ over past iterations to reduce the bias. The use of cFIM is not ideal because it may result in slow convergence, but it is the best available option. The *standard* Fisher information matrix, $\mathbf{I}_{\text{FIM}}(\Theta) = -E_{\mathbf{Y}}[\nabla_{\Theta}^2 L_s(\Theta; \mathbf{Y})]$, is seldom explicit and thus cannot be used as a preconditioner. However, we do need to estimate either the standard or the *observed* Fisher information matrix, oFIM, $\mathbf{I}_{\text{oFIM}}(\Theta) = -\nabla_{\Theta}^2 L_s(\Theta; \mathbf{Y})$, to calculate confidence intervals for the estimated parameters. We estimate $\mathbf{I}_{\text{oFIM}}(\Theta)$ by using

Louis's identity (Louis, 1982),

$$\mathbf{I}_{\text{OFIM}}(\boldsymbol{\Theta}) = -E_{\mathbf{V}}[\nabla_{\boldsymbol{\Theta}}^2 L_s(\boldsymbol{\Theta}; \mathbf{Y}, \mathbf{V}) | \mathbf{Y}] - \text{var}_{\mathbf{V}}\{\nabla_{\boldsymbol{\Theta}} L_s(\boldsymbol{\Theta}; \mathbf{Y}, \mathbf{V}) | \mathbf{Y}\}. \quad (20)$$

Both terms on the right-hand side of equation (20) can be estimated by MC sampling, as proposed for $\nabla_{\boldsymbol{\Theta}} L_s(\boldsymbol{\Theta}; \mathbf{Y})$ in equation (17). We could also estimate $\mathbf{I}_{\text{FIM}}(\boldsymbol{\Theta})$ by an additional sampling step, using the fact that $\mathbf{I}_{\text{FIM}}(\boldsymbol{\Theta}) = E_{\mathbf{Y}}[\mathbf{I}_{\text{OFIM}}(\boldsymbol{\Theta})]$. For more details on the calculation of the gradients required and oFIM, see Appendix A.

4.3. Multiple-chain estimation

A drawback of the estimation procedure that was described in the previous subsection is that the stochastic nature of the method can make it difficult to determine a suitable stopping criterion. To overcome this problem, we propose running several independent estimation procedures in parallel, starting from the same initial guess, i.e., using algorithm (18), we compute N_r different estimates of $\boldsymbol{\Theta}^{(n)}$, $\{\boldsymbol{\Theta}_i^{(n)} : i = 1, \dots, N_r\}$, in parallel started from the same initial guess and using independent stochastic estimates of the gradient and preconditioner.

We combine these estimates by taking the mean of the N_r -estimates and calculate the estimate of the corresponding MC standard deviation $\sigma^{(n)}$. These statistics are calculated, for the j th parameter, as

$$\Theta_j^{(n)} = \frac{1}{N_r} \sum_{i=1}^{N_r} \Theta_{i,j}^{(n)},$$

and

$$\sigma^{(n)} = \left\{ \frac{1}{N_r} \sum_{i=1}^{N_r} (\Theta_{i,j}^{(n)} - \Theta_j^{(n)})^2 \right\}^{1/2},$$

where $\Theta_{i,j}^{(n)}$ denotes the j th element in $\boldsymbol{\Theta}_i^{(n)}$. We run each chain in batches. Whenever $n = kN_b$, for $k = 4, 5, 6, \dots$ being the number of batches and N_b a chosen batch size (we use 1000 as default), we check whether each parameter Θ_j , the j th element of $\boldsymbol{\Theta}$, has converged on the basis of two criteria. The first is that $\sigma_j^{(n)} / \Theta_j^{(n)}$ should be smaller than a threshold, e.g. 0.1. This means that the MC variance should be sufficiently small for each parameter. The second convergence criterion is that the rate of change for each parameter should be sufficiently small. To check this, we estimate the intercept and slope of a simple linear regression model fit to the last four batch estimates of Θ_j (i.e. the values $\Theta_j^{(k-l)N_b}$ for $l = 0, \dots, 3$) as the outcome and iteration as the input. We then check whether the magnitude of the slope is not significantly larger than some constant (such as 0.01) times the magnitude of the intercept. We conclude convergence of a parameter if both criteria are satisfied and stop the estimation procedure if all parameters converge. An example of the trajectories for the multiple-chain estimation procedure is shown in Fig. 3.

The use of multiple chains has extra advantages, in addition to providing stopping criteria. The first is that the combined estimates that are based on multiple chains naturally have lower MC variances than the estimates that are based on using a single chain. The second is that the estimates of the MC variances can be used when computing confidence intervals for the parameters, which can improve coverages of Wald-type confidence intervals. Specifically, to compute a confidence interval for Θ_j , we use $\Theta_j \pm z_{\alpha/2} \sqrt{\{(\sigma_j^{(n)})^2 + \sigma_j^2\}}$, where $\sigma_j^{(n)}$ is the estimate of the MC standard deviation for the final iteration n , and σ_j^2 denotes the j th diagonal element of the inverse observed Fisher information matrix. Details on the calculation of the Fisher information matrix are given in Appendix A.

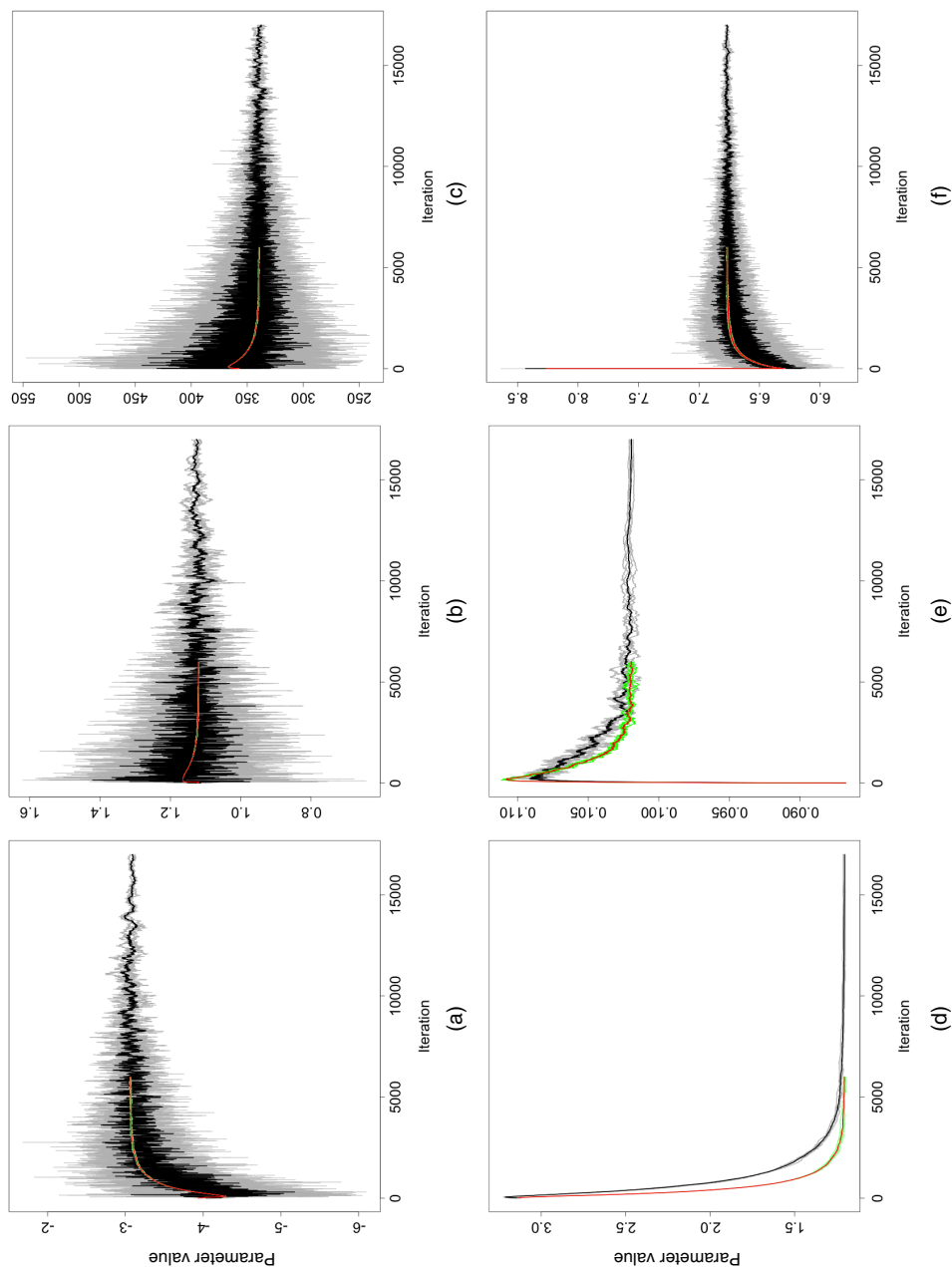


Fig. 3. Stochastic gradient estimation paths for results without subsampling (—) and with subsampling 20% of the patients in each iteration (—) (—), the four individual paths that are run in parallel to compute these curves for results with and without subsampling; the automatic stopping criterion caused the estimation without subsampling to terminate after 6000 iterations whereas 17000 iterations were used with subsampling): (a) fixed effect; (b) random effect; (c) operator κ ; (d) operator τ ; (e) noise σ

4.4. Subsampling with fixed effects: the grouped subsampler

Another issue with the subsampling method that was described in Section 4.2 is that the subsampled matrices of covariates, \mathbf{x}_i , may not be of full rank. If this is so, none of the preconditioners that were described above can be used. In contrast, regular subsampling without any preconditioners may result in large MC variation in the estimated gradient. The cystic fibrosis case-study that we shall describe in Section 6.1 provides an example. These data are stratified into birth cohorts whose effects are important, but one of the cohorts contains only seven patients. This issue is related to subsampling for S -estimation algorithms in linear regression models (Koller and Stahel, 2016), but we could not find a satisfactory solution in the literature that could be applied in the current context. To address the issue, we therefore introduce the following subsampling procedure, which we call the *grouped subsampler*. The procedure first builds $k + 1$ groups of subjects, $\mathcal{G}_0, \mathcal{G}_1, \dots, \mathcal{G}_k$, in such a way that the matrices $\mathbf{x}_{\mathcal{G}_j} = \sum_{i \in \mathcal{G}_j} \mathbf{x}_i \mathbf{x}_i^T$ have full rank for $j \geq 1$.

The groups are built iteratively starting with \mathcal{G}_1 . To this group the first subject is added and one checks whether the covariate matrix $\mathbf{x}_{\mathcal{G}_1} = \mathbf{x}_1 \mathbf{x}_1^T$ has full rank. If this is so, the formation of \mathcal{G}_1 is complete. Otherwise more subjects need to be added to the group. If $\mathbf{x}_{\mathcal{G}_1} + \mathbf{x}_2 \mathbf{x}_2^T$ has a larger rank than $\mathbf{x}_{\mathcal{G}_1}$, the second subject is added to the group and $\mathbf{x}_{\mathcal{G}_1}$ is updated to $\mathbf{x}_{\mathcal{G}_1} + \mathbf{x}_2 \mathbf{x}_2^T$. At this stage, the formation of \mathcal{G}_1 is complete if $\mathbf{x}_{\mathcal{G}_1}$ has full column rank; otherwise the procedure continues by adding more subjects in order until $\mathbf{x}_{\mathcal{G}_1}$ has full rank or until no further subjects are left. If the formation of \mathcal{G}_1 terminates because of a lack of further subjects, one cannot estimate the model on the basis of the available subjects. Otherwise, further groups $\mathcal{G}_2, \mathcal{G}_3, \dots$ are constructed iteratively in the same way: subjects, who are not in any of the previous groups, are added to the group \mathcal{G}_k until the covariate matrix $\mathbf{x}_{\mathcal{G}_k}$ has full rank. At some point, the group formation will terminate because of the lack of further subjects. If this happens during the formation of a group \mathcal{G}_k , this group is removed since its covariate matrix does not have full rank. Finally, any subjects who have not been assigned to a group are placed in the group \mathcal{G}_0 . The procedure for forming the groups is described in pseudocode in algorithm 1 in Appendix A.4.

The groups are created as an initial stage before the estimation procedure begins. During the estimation procedure a subsampling step is performed as follows. Assume that we want to sample a proportion $p \in (0, 1)$ of the subjects. Let n_g be the total number of subjects in the groups $\mathcal{G}_0, \dots, \mathcal{G}_k$. To ensure that we obtain a sample for which the subsampled covariate matrix has full column rank, we need to sample all subjects from at least one of the groups $\mathcal{G}_1, \dots, \mathcal{G}_k$. Given this restriction, we would like to sample approximately pn_g subjects from the groups $\mathcal{G}_1, \dots, \mathcal{G}_k$ as well as pn_0 subjects from \mathcal{G}_0 . To do so, we first sample all the subjects from $m_g = \lceil pk \rceil$ out of the groups $\mathcal{G}_1, \dots, \mathcal{G}_k$ chosen at random; then we sample $m_0 = \min\{\max(\lceil pN - Mk/n_g \rceil, 1), n_0\}$ subjects at random from \mathcal{G}_0 ; here, $\lceil \cdot \rceil$ denotes the function that outputs the nearest integer.

To obtain an unbiased estimate of the gradient in the estimation step when using the subsampling procedure, we assign weights k/m_g to the subjects sampled from the groups $\mathcal{G}_g, g = 1, \dots, k$, and weight n_0/m_0 to those sampled from \mathcal{G}_0 , so that each subject has weight 1 divided by the probability of their being sampled. The fact that we can obtain unbiased estimates by using the grouped subsampler is crucial. Many apparently natural subsampling solutions to the column rank problem, e.g. the solution that samples subjects until the matrix has full column rank, will not work because they produce samples that cannot easily be weighted to obtain unbiased estimates.

4.5. Prediction

Suppose that, for a given subject i , we want to predict the value of the latent process at a given time t_k , i.e. $Y_{ik}^* = \mathbf{x}_{ik}^T \boldsymbol{\beta} + \mathbf{d}_{ik}^T \mathbf{U}_i + W_i(t_k)$. Different types of predictions may be defined depending

on the scientific interests of a specific application. The first is smoothing prediction, where the quantity of interest is the value given all available data for that patient, $Y_{ik}^*|\mathbf{Y}_i$. The second is filtering, where the quantity of interest is the value given all data collected before the time t_k , $Y_{ik}^*|\mathbf{Y}_i^{k,f}$, where $\mathbf{Y}_i^{k,f} = \{Y_{ij} : t_{ij} < t_k\}$. The third is nowcasting, where the quantity of interest is the value given all data collected up to and including the time t_k , $Y_{ik}^*|\mathbf{Y}_i^{k,n}$, where $\mathbf{Y}_i^{k,n} = \{Y_{ij} : t_{ij} \leq t_k\}$.

For all three cases, we can sample from the relevant predictive distribution, using the same Gibbs sampler as was used when estimating the gradient (just conditioning on different sets of data). The details of this Gibbs sampler are given in Appendix A.1. Given M values obtained from the Gibbs sampler, we approximate the quantities of interest such as the predictive mean or the quantiles of the predictive distribution needed for constructing prediction intervals, using MC integration.

5. Model validation and model selection

An obvious question is how to decide when a non-Gaussian model is preferable to the standard Gaussian model. A natural first step would be to check the validity of a Gaussian model through its marginal residuals, $Y_{ij} - \mathbf{x}_{ij}^T \beta$, standardized by their variances. These standardized residuals can be plotted against theoretical quantiles of the standard normal distribution to check for departures from normality. However, this plot would not show which of the model components are the source of the non-Gaussianity. Also, the approach cannot easily be used to check the validity of any given non-Gaussian model, since we do not necessarily know the true distribution of its residuals.

We therefore check the validity of a model, Gaussian or not, by predicting each model component from the data given the estimated parameters and compare the distribution of each component with the corresponding quantity for data simulated from the model. Thus, to check the validity of the error model, we compute $\hat{Z}_{ij} = Y_{ij} - \mathbf{x}_{ij}^T \beta - E[U_i + W_i(t_{ij})|\mathbf{Y}_i]$, then simulate a new data set $\mathbf{Y}_i^s = \{\mathbf{Y}_{ij}^s : j = 1, \dots, n_i\}$ for each subject according to the model and compute $\hat{Z}_{ij}^s = \mathbf{Y}_{ij}^s - \mathbf{x}_{ij}^T \beta - E[U_i + W_i(t_{ij})|\mathbf{Y}_i^s]$ based on the simulated values. We can then visualize the model fit by using, for example, Q - Q -plots between \hat{Z}_{ij} and \hat{Z}_{ij}^s . If these plots deviate from the line of equality, this indicates that the model does not fit the data. For a more formal assessment, we can repeat the procedure for K different simulated data sets and compute a joint simulation envelope. If the envelope does not contain the line of equality we can reject that the data are generated by the model.

Using the same simulations, we can similarly assess the fit of the random effects U_i , comparing the quantiles of $\hat{U}_i = E[U_i|\mathbf{Y}_i]$ and $\hat{U}_i^s = E[U_i|\mathbf{Y}_i^s]$. To assess the fit of the process $W_i(t)$, we apply a similar procedure to the innovations of the process dL_i . Using the innovations, rather than the $W_i(t)$, avoids the correlation-induced distortion of the empirical marginal distributions of $W_i(t)$.

6. Case-studies

6.1. Natural progression of lung function in cystic fibrosis patients

Our first application uses data on the lung function of cystic fibrosis patients, taken from the Danish cystic fibrosis register. The patients are all aged over 5 years and entered the database between 1969 and 2010. The outcome variable is %FEV1 (percentage predicted forced expiratory volume in 1 s), which is a measure of lung function that is widely used as a descriptor of severity of disease (Davies and Alton, 2009). The data, previously analysed by Taylor-Robinson

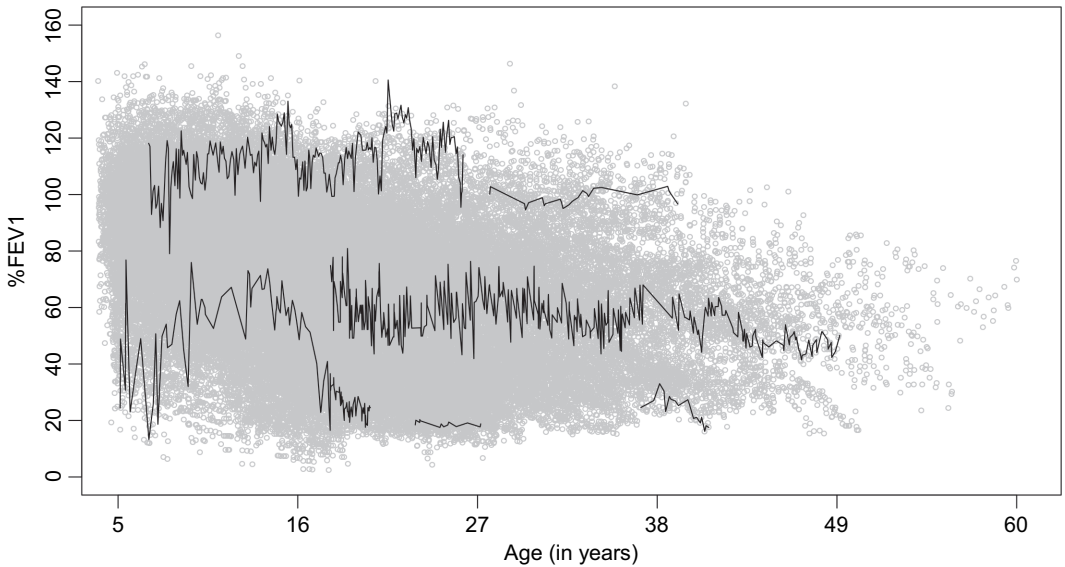


Fig. 4. %FEV1 measurements against age (in years) in the background as a grey scatter plot: —, data on six patients

et al. (2012), contain 70448 measurements of %FEV1 on 479 patients with follow-up times approximately 1 month apart. For the analysis that is reported here, three patients who provided only one %FEV1-measurement have been excluded. Hence, 476 patients are available for the current analysis. Available covariates are sex, age, birth cohort (decadal), presence or absence of pancreatic sufficiency, presence or absence of *diabetes mellitus* and the number of years with pseudomonas: a bacterial infection to which cystic fibrosis patients are susceptible. The number of repeated measures per patient ranges between 2 and 597 with a median of 101.5. Total follow-up times range between 0.1 and 31.5 years with a median of 10.5. Of the 476 patients, 233 (48.9%) are female, 20 (4.2%) have pancreatic sufficiency and 14 (2.9%) have diabetes. Baseline ages range between 5.0 and 48.1 years with a median of 7.0 years. Cohort numbers are 7 (1.5%), 42 (8.8%), 109 (22.9%), 105 (22.1%), 141 (29.6%) and 72 (15.1%) for birth cohorts of 1948–1957, 1958–1967, ..., 1998–2007 respectively. Baseline %FEV1 values range between 10.4 and 140.3 with a mean of 78.5. Fig. 4 shows traces for six patients, chosen to illustrate a range of total follow-up times and patterns of the outcome variable %FEV1.

Fitting a model to these data serves two purposes. The first is to characterize the mean response profile of lung function in cystic fibrosis patients, adjusted for relevant covariates. The second is to quantify the extent to which a subject's early results are predictive of their long-term prognosis.

We consider %FEV1 as the outcome, Y , and specify mixed effects models that fall within the general framework of equation (9). Specifically, we consider

$$Y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + U_i + W_i(t_{ij}) + Z_{ij}, \quad (21)$$

where each \mathbf{x}_{ij} contains a number of explanatory variables, as listed in Table 3. We model $W_i(t)$ as the solution to the stochastic differential equation $(\kappa^2 - d^2/dt^2)W_i(t) = dL_i(t)$, which implies that $W_i(t)$ has a Matérn covariance function with smoothness parameter $\frac{3}{2}$. We also considered

Table 3. Estimates of the fixed effects for the normal and NIG models†

	Results for the normal model			Results for the NIG model		
	Estimate	Standard error	p-value	Estimate	Standard error	p-value
Intercept	68.23	2.07	< 0.001	72.98	2.06	< 0.001
Diabetes	−3.05	0.52	< 0.001	−1.76	0.44	< 0.001
Years after pseudomonas infection	−0.46	0.08	< 0.001	−0.45	0.07	< 0.001
Age	−0.25	0.08	0.002	−0.47	0.07	< 0.001
cohort 1948	1.71	10.09	0.865	2.83	9.54	0.767
cohort 1958	−4.05	3.97	0.308	−8.12	3.82	0.034
cohort 1978	18.21	2.93	< 0.001	14.5	2.87	< 0.001
cohort 1988	26.37	2.84	< 0.001	22.52	2.74	< 0.001
cohort 1998	28.95	3.59	< 0.001	23.28	3.34	< 0.001
Pancreatic sufficiency	0.98	6.26	0.876	4.26	5.71	0.456
Age * cohort 1948	−0.08	0.19	0.674	−0.06	0.17	0.724
Age * cohort 1958	0.09	0.08	0.261	0.28	0.07	< 0.001
Age * cohort 1978	−0.83	0.09	< 0.001	−0.71	0.09	< 0.001
Age * cohort 1988	−0.79	0.15	< 0.001	−0.81	0.14	< 0.001
Age * cohort 1998	0.49	0.54	0.364	0.35	0.43	0.416
Age * pancreatic sufficiency	1.12	0.3	< 0.001	0.81	0.27	0.003

†Age is centred at 5. Cohort 1968, absence of diabetes and absence of pancreatic sufficiency are the reference categories.

a model with an exponential covariance function, as in Taylor-Robinson *et al.* (2012), but this gave a worse fit to the data.

In this example, cohort effects are substantial, reflecting general improvements in the treatments that are available to cystic fibrosis patients over the time period that is concerned. This, coupled with the small numbers of patients in some cohorts (e.g. seven patients in 1948–1957), explains why the grouped subsampler that was described in Section 4.4 is needed.

To illustrate the effect of the subsampling using the proposed grouped subsampler, we first fit a Gaussian model, i.e. assuming Gaussian distributions for U_i , $W_i(t)$ and Z_{ij} , with and without subsampling. In the with-subsampling case, we subsample 20% of the patients, i.e. 96 out of 476. The resulting parameter tracks of the optimizer (for six of the model parameters) can be seen in Fig. 3. In this example, there are $k = 7$ subsampling groups, with an average group size of eight subjects, and we sampled two groups at each iteration. The running time for each iteration scales linearly with M , the number of patients who are subsampled at each iteration. Thus, subsampling reduces computing time per iteration by a factor of almost 5 in this case. However, since 17000 iterations were needed to meet the convergence criteria with subsampling, compared with 6000 iterations without, the total computation time was reduced by a factor of 1.75. The variances of the subsampled estimates are slightly larger, but the final parameter estimates are almost identical. For applications with data for many more subjects, e.g. as in the renal case-study to be presented in the next section, the computational gain by subsampling would be larger.

To assess the suitability of the Gaussian distributional assumption, we inspected Q – Q -plots of the standardized marginal residuals, $Y_{ij} - \mathbf{x}_{ij}^T \beta$. Fig. 5 suggests some departure from the Gaussian distribution but, as each marginal residual is composed of U_i , $W_i(\cdot)$ and Z_{ij} , the Q – Q -plot cannot detect the source of the departure. We therefore also look at the Q – Q -plots for

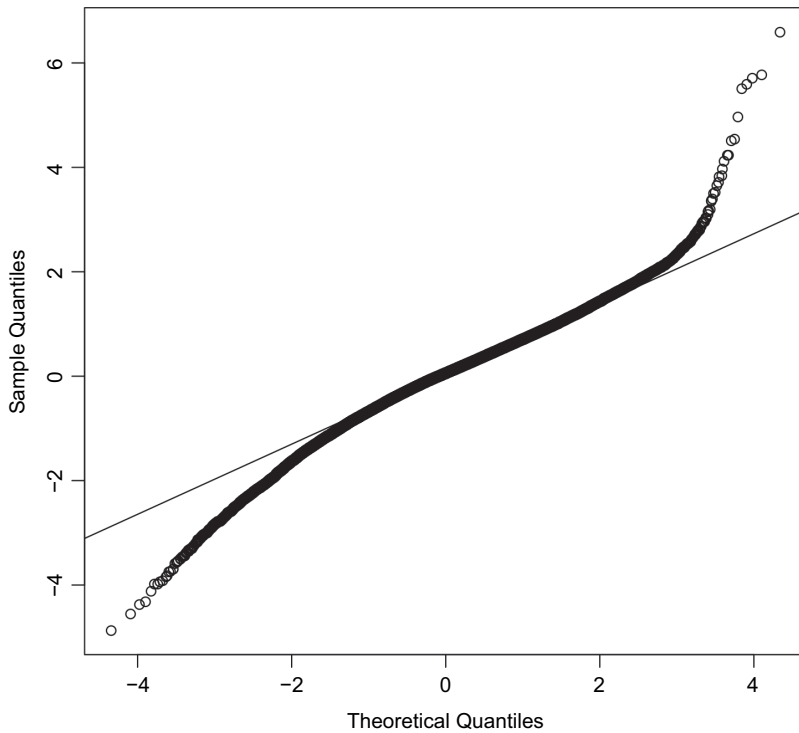


Fig. 5. Q - Q -plot of standardized residuals under the Gaussian assumption for the cystic fibrosis example

the various model components, as explained in Section 5. The results for the Gaussian model can be seen in Figs 6(a)–6(c). The normality assumption seems to be valid for the random effects U_i , but both the process, $W_i(\cdot)$, and the errors, Z_{ij} , seem to have departures from the normal distribution. On the basis of these, we fitted a model with the NIG assumption for $W_i(\cdot)$ and Z_{ij} and normal distribution assumption for U_i . The corresponding Q - Q -plots for this model form Figs 6(a)–6(f). All three components now fit fairly well, although the random effects for the data seem to have a slightly skewed distribution. Thus, as a final model we fitted a model with an NIG assumption for each of the three model components. Figs 6(g)–6(i) show the Q - Q -plots for this model. The NIG assumption for the random effects improves the fit that the distributional assumptions are now reasonably good for all three components. We conclude that the NIG assumption for all three stochastic components is the most appropriate model for these data.

The estimates of the fixed effect parameters β for the normal and NIG models are shown in Table 3. Standard error estimates, obtained by using the observed Fisher matrix, are generally lower under the NIG than under the Gaussian assumption. With regard to statistical significance, p -values indicate the same judgement on significance, except for pancreatic sufficiency, main effect of cohort 1958–1967 and its interaction with age. Note that both pancreatic sufficiency and cohort 1958–1967 are highly unbalanced with only 20 positive out of 476 and only 42 subjects in the 1958–1967 cohort.

In Section 7.1, we report the results from a simulation study to validate these findings. Two important conclusions based on the simulation study are that

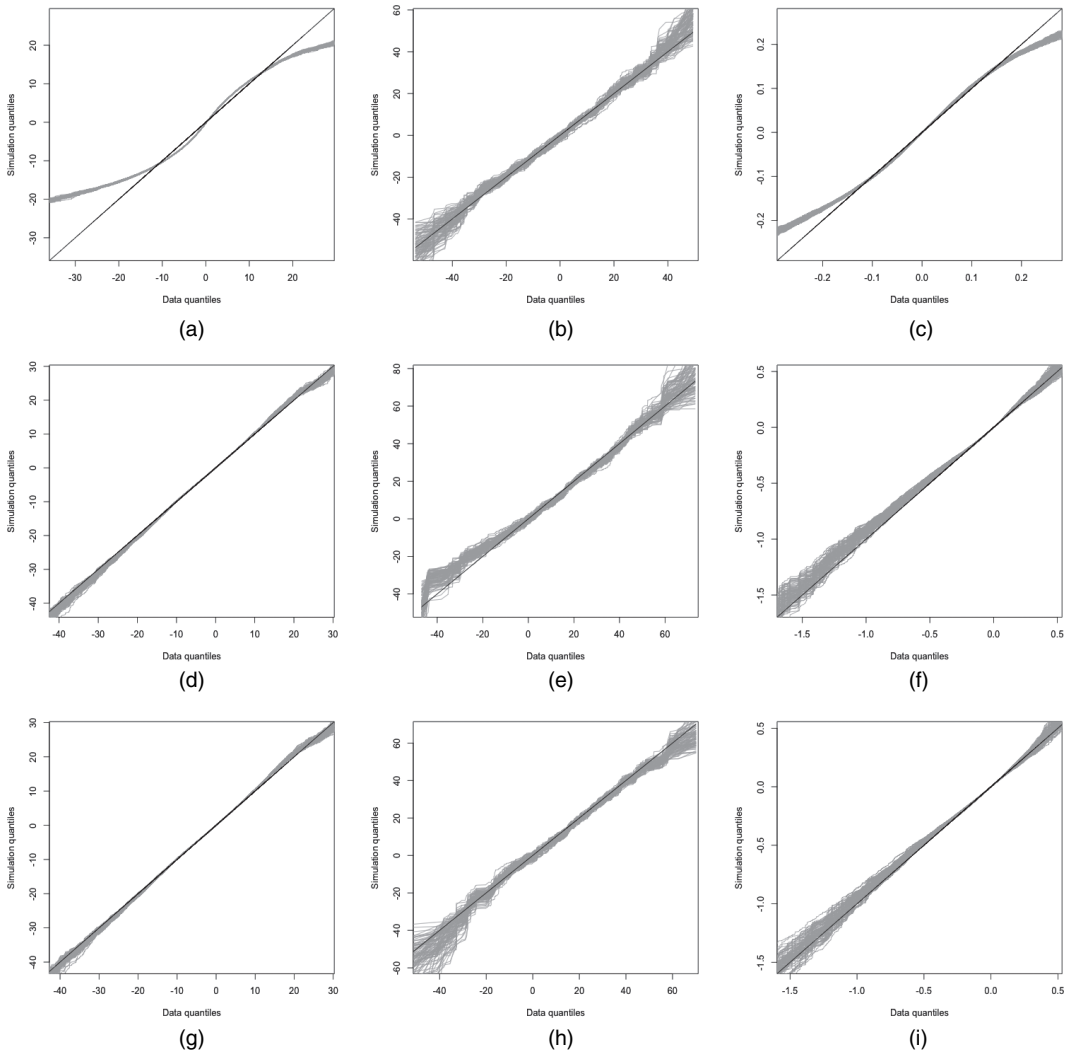


Fig. 6. Q - Q -plots for the components of the models in Section 6.1 for the cystic fibrosis example, based on 100 simulated data sets from each model: for the model with Gaussian (a) Z_{ij} , (b) U_i and (c) $W_i(t)$; for the model with (d) NIG-distributed Z_{ij} , (e) Gaussian U_i and (f) NIG-distributed $W_i(t)$; for the model with NIG distributions for (g) Z_{ij} , (h) U_i and (i) $W_i(t)$

- (a) it is important to include a random-process term $W_i(t)$ in the model to obtain reliable inferences for the fixed effects and
- (b) a Gaussian assumption for $W_i(t)$ may still deliver reliable inferences regarding fixed effects, when the data show signs of non-Gaussianity.

6.2. Progression towards end stage renal failure

Our second application uses clinical data on kidney function of primary care patients from the northern English city of Salford who are in high-risk groups for chronic kidney disease. The outcome variable is eGFR (estimated glomerular filtration rate, in millilitres per minute per

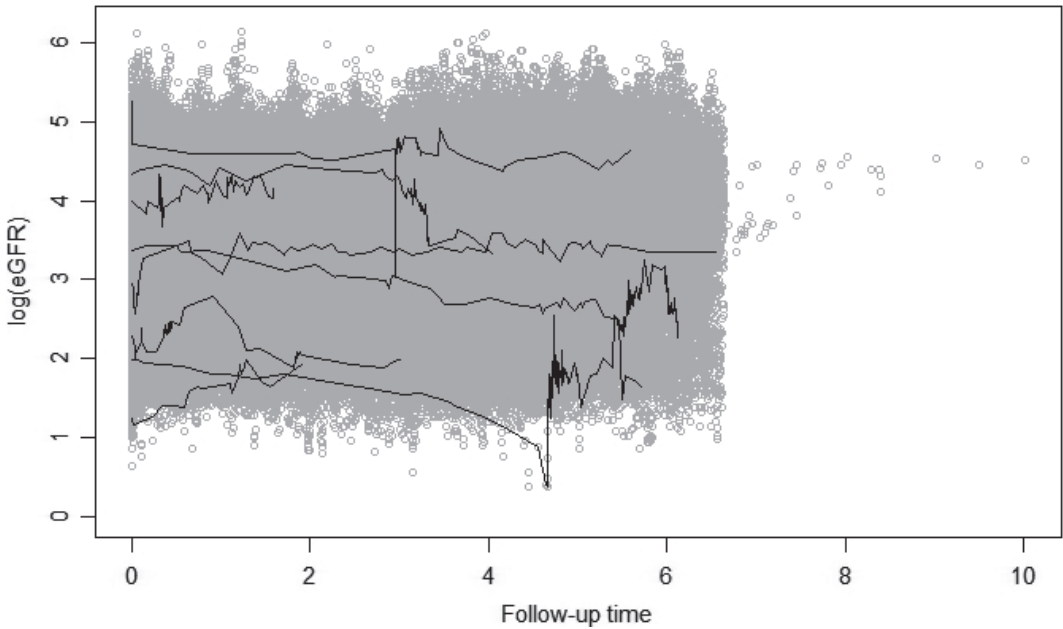


Fig. 7. eGFR-measurements (on a log-scale) against follow-up time (in years) in the background as a scatter plot: data on eight patients are highlighted by black lines connecting successive measurements

1.73 metres squared of body surface area): a proxy measurement for the patient's renal function calculated as

$$\text{eGFR} = 175 \left(\frac{\text{SCr}}{88.4} \right)^{-1.154} \text{age}^{-0.203} \times 0.742^{I(\text{female})} \times 1.21^{I(\text{black})}, \quad (22)$$

where SCr stands for serum creatinine measured in micromoles per litre (Levey *et al.*, 1999).

The data, which were previously analysed by Diggle *et al.* (2015), contain a total of 392870 measurements on 22910 patients, for whom the total follow-up time ranged from 0 (i.e. only baseline data are available) to 10.0 years, whereas the number of measurements of eGFR ranged from 1 to 305. Among the 22910 patients, 11833 (51.7%) were male. Baseline ages ranged between 13.7 and 102.1 with a mean of 65.4.

Fig. 7 shows traces for eight patients, chosen to illustrate some particularly challenging features of the data. The unusually high degree of irregularity in the follow-up times reflects the fact that the data derive from routine clinical practice. In particular, some patients provided many repeated measurements over a relatively short time period, probably during episodes of intercurrent illness.

Clinical care guidelines in the UK include a recommendation that any person in primary care who appears to be losing kidney function at a relative rate of at least 5% per year should be considered for referral to specialist secondary care. Our primary objective in analysing these data is therefore to develop a method for identifying, for each subject and in realtime, when this criterion is first met.

As in Diggle *et al.* (2015), we use a log-transformed outcome variable $Y = \log(\text{eGFR})$ and specify a model of the form

$$Y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + U_i + W_i(t_{ij}) + Z_{ij}. \quad (23)$$

In model (23), each \mathbf{x}_{ij} includes sex, baseline age, follow-up time t_{ij} and a piecewise linear function of age with a slope change at age 56.5 years. The processes $W_i(t)$ are integrated random walks as in Diggle *et al.* (2015).

As for the previous example, we first fit the model under Gaussian assumptions for the U_i -, $W_i(\cdot)$ - and Z_{ij} -components. The Q - Q -plot for the standardized residuals that is shown as Fig. 5 of Diggle *et al.* (2015) clearly indicates longer-than-Gaussian tails. As for the cystic fibrosis case-study, we compute the Q - Q -plots for each model component (Fig. 8), which suggest that the Gaussian assumption is not valid for any of the model components. Therefore, we proceed by assuming NIG distributions for each of the three stochastic components. The fit is much improved, albeit with some discrepancy between the data and model in the lower tail of the U_i and the upper tail of the $W_i(t)$.

Fig. 9 shows, for two patients, their observed data and the concurrent ('nowcasting') probabilities of meeting the clinical guideline for referral to specialist care. Results are shown for the Gaussian and NIG models. As would be expected, for each patient the general pattern of the predictive probabilities is similar under both modelling assumptions, but there are some substantial quantitative differences and the ranking of each pair of predictive probabilities is not consistent. The two sets of model-based predictions reflect different partitionings of the inpatient variation into signal and noise components, and the balance between the two is affected in subtle ways by the pattern of follow-up times and their associated measurements.

To put these differences in context, in Section 7.2 we report the results of a simulation study, where we find that the distributional assumptions have a strong effect on the predictive performance.

7. Simulation studies

7.1. Fixed effects estimation

We conduct a simulation study to investigate the extent to which distributional assumptions affect the validity and/or efficiency of estimators for $\boldsymbol{\beta}$. We focus on evaluating the bias and coverage properties of the estimators by using the models for the cystic fibrosis patients that were presented in Section 6.1, but with a reduction in size to 256 patients covering the cohorts between 1958 and 1978 to reduce computation time.

We consider two simulation models, denoted by normal and NIG. In the first, all three stochastic components are Gaussian, with parameters set at the estimates that were obtained from the cystic fibrosis data, whereas in the second all three stochastic components are NIG distributed, again with parameters set at their estimates from the cystic fibrosis data; see Table 3 for $\boldsymbol{\beta}$ and Table 4 for the parameters of U_i , $W_i(t)$ and Z_{ij} . We generate 250 replicate data sets from each of the two simulation models. For each data set, we fit both simulation models and three additional 'wrong' Gaussian models: a standard multiple linear regression model, a random-intercept-only model and a random-intercept and random-slope model. In each case, we evaluate the empirical bias of each parameter estimator and the coverage of nominal 90% confidence intervals over the 250 replicates. The confidence intervals are computed as explained in Section 4.3. The results are presented in Tables 5 and 6. Important findings are as follows.

- (a) For both the normal and the NIG simulation models, the linear regression model and the two random-effects models give very poor coverages, indicating that inclusion of the process component $W_i(t)$ is crucial for making correct inferences about the fixed effects.

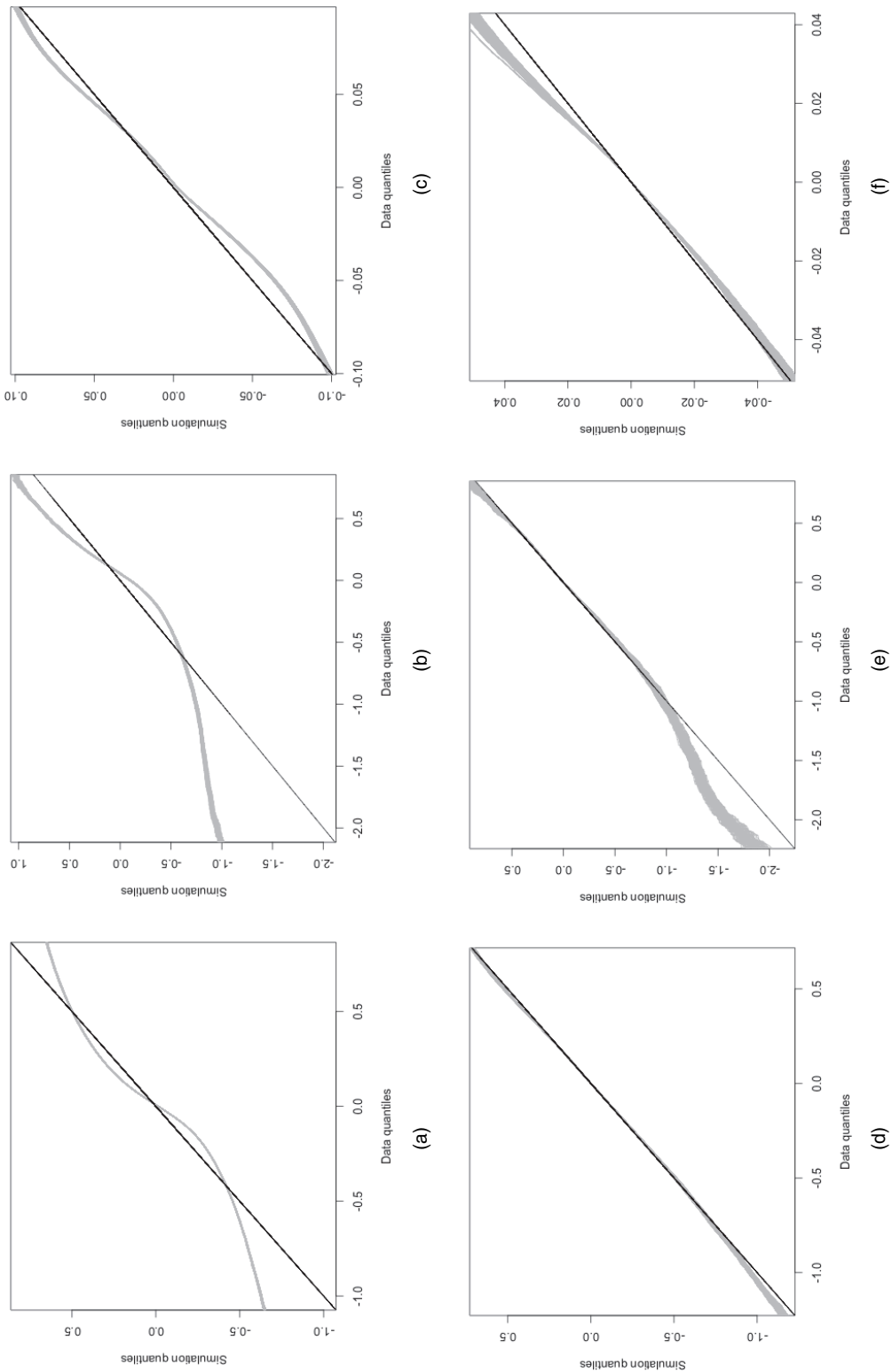


Fig. 8. Q - Q -plots for the three model components of the (a)–(c) Gaussian and (d)–(f) NIG models for the renal failure data: (a) normal $W_{ij}(t)$; (b) normal Z_{ij} ; (c) normal U_{ij} ; (d) NIG $W_{ij}(t)$; (e) NIG Z_{ij} ; (f) NIG U_{ij} .

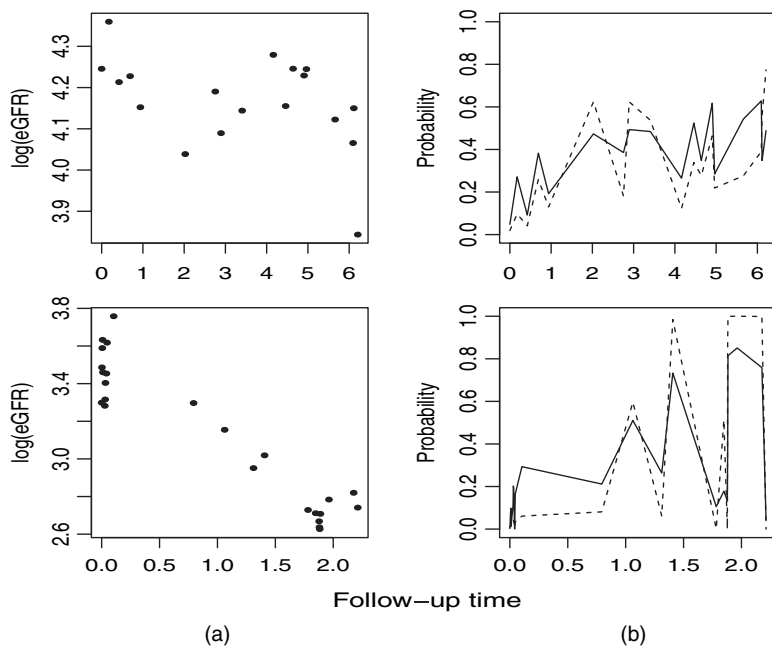


Fig. 9. (a) Follow-up time (in years) *versus* log(eGFR) for two patients and (b) probabilities of meeting the clinical guideline for the patients (—, normal distribution; - - -, NIG distribution)

Table 4. Parameters for U_i , $W_i(t)$ and Z_{ij} used in the simulation study

<i>Model</i>	Σ^U	μ^U	ν^U	σ^Z	μ^Z	ν^Z	τ	κ	μ^W	ν^W
Normal	341.59			6.76			0.10	1.22		
NIG	294.36	-47.24	19.26	6.59	-1.16	0.52	0.089	1.21	-0.08	0.34

- (b) For the normal simulation model, the performances of the NIG model and the normal model are similar, indicating that the NIG model gives reliable estimates even if the data are Gaussian.
- (c) For the NIG simulation model, the confidence intervals for the normal model have almost the correct coverage. The NIG model has similar coverage and bias to those for the normal model. However, the standard deviations of the estimates for the NIG model are smaller in the case of NIG data, which makes the confidence intervals tighter and thus indicates that we obtain a higher efficiency when using the NIG model when the data are non-Gaussian.

The overall conclusion from this small simulation study is that it is important to include a random process in the model to obtain reliable inference of the fixed effects but that, for this purpose, a Gaussian model can give reliable inferences even if the data show signs of being non-Gaussian.

7.2. Prediction accuracy

To study the importance of the distributional assumptions on prediction, we perform a simulation study based on the renal failure application, presented in Section 6.2. We simulate new

Table 5. Results for the simulation study on fixed effects estimation†

	<i>Results for the linear model</i>			<i>Results for the random model</i>			<i>Results for the random- slope model</i>			<i>Results for the normal model</i>			<i>Results for the NIG model</i>		
	<i>Bi</i>	<i>C</i>	<i>Se</i>	<i>Bi</i>	<i>C</i>	<i>Se</i>	<i>Bi</i>	<i>C</i>	<i>Se</i>	<i>Bi</i>	<i>C</i>	<i>Se</i>	<i>Bi</i>	<i>C</i>	<i>Se</i>
Intercept	−1	15	0.28	−1	84	1.89	−1	82	1.99	−1	90	2.04	−1	85	1.99
Diabetes	−4	8	0.27	−1	30	0.22	−1	24	0.27	2	90	0.58	2	88	0.58
Years after pseudomonas infection	2	11	0.02	1	25	0.02	1	77	0.08	0	90	0.08	−1	88	0.08
Age	−3	14	0.02	−2	25	0.02	−1	27	0.03	0	92	0.08	1	92	0.08
cohort 1958	−7	16	0.62	6	82	3.56	9	73	3.7	4	88	3.91	3	86	3.79
cohort 1978	2	16	0.47	2	85	2.67	2	79	2.77	2	89	2.89	2	85	2.79
Pancreatic sufficiency	−17	17	2.91	−13	65	6.92	−10	67	7.16	−3	93	10.43	−3	93	10.42
Age * cohort 1958	−14	16	0.03	8	20	0.02	13	33	0.05	5	89	0.08	5	88	0.08
Age * cohort 1978	1	19	0.03	0	30	0.02	0	29	0.04	0	95	0.09	0	96	0.09
Age * pancreatic sufficiency	2	20	0.16	3	36	0.16	3	34	0.16	2	94	0.42	2	92	0.43

†250 data sets are generated from the normal model. Bi, percentage relative bias (the bias divided by the true parameter times 100); C, percentages of the 90% confidence intervals that cover the true value; Se, standard deviations of the estimates.

Table 6. Results for the simulation study on fixed effects estimation†

	<i>Results for the linear model</i>			<i>Results for the random model</i>			<i>Results for the random- slope model</i>			<i>Results for the normal model</i>			<i>Results for the NIG model</i>		
	<i>Bi</i>	<i>C</i>	<i>Se</i>	<i>Bi</i>	<i>C</i>	<i>Se</i>	<i>Bi</i>	<i>C</i>	<i>Se</i>	<i>Bi</i>	<i>C</i>	<i>Se</i>	<i>Bi</i>	<i>C</i>	<i>Se</i>
Intercept	0	10	0.31	0	85	2.08	0	83	2.22	0	92	2.26	0	90	2.16
Diabetes	−1	8	0.3	−3	24	0.24	−9	27	0.3	−1	89	0.62	0	90	0.47
Years after pseudomonas infection	2	7	0.02	−3	27	0.02	−2	73	0.09	−1	92	0.09	0	89	0.07
Age	−2	13	0.02	3	25	0.02	2	32	0.03	1	90	0.09	1	86	0.07
cohort 1958	0	13	0.68	−2	84	3.91	−2	71	4.1	−2	88	4.34	−2	87	3.99
cohort 1978	2	13	0.52	1	83	2.94	2	82	3.07	2	88	3.2	2	88	2.96
Pancreatic sufficiency	−18	21	3.21	16	61	7.62	15	62	7.98	11	89	11.75	9	86	10.25
Age * cohort 1958	1	14	0.03	1	22	0.02	−2	27	0.06	−1	94	0.09	−1	90	0.07
Age * cohort 1978	1	27	0.04	0	27	0.02	1	24	0.04	0	87	0.1	1	86	0.08
Age * PS	3	17	0.18	−3	35	0.18	−3	32	0.17	−2	88	0.49	−1	90	0.4

†250 data sets are generated from the NIG model. For the explanations, see the caption of Table 5.

data from the fitted NIG model, for the two patients who are shown in Fig. 9. For each simulated sequence $Y_{ij} = \log(\text{eGFR})$ at follow-up times t_{ij} , we then use the fitted NIG and Gaussian models that were reported in Section 6.2 to obtain the nowcasting predictions. Table 7 shows the results based on 100 simulated data sets, using four summaries of predictive performance: the mean absolute error, MAE, root-mean-squared error, RMSE, mean coverage of 95% prediction intervals, COV, and the average width of these prediction intervals, Width. For the two models, we also look at the model-based predictive probabilities of meeting the clinical guideline for referral to specialist care, and we compare these with the corresponding simulated values of

Table 7. Results for the simulation study for prediction†

<i>Model</i>	<i>MAE</i> ($\times 100$)	<i>RMSE</i> ($\times 100$)	<i>COV</i>	<i>Width</i>	<i>AUC</i>
Normal	9.122	2.705	0.912	0.402	0.959
NIG	5.849	0.838	0.935	0.283	0.995

†The results are based on 100 simulated data sets for the two patients for whom results are presented in Fig. 9.

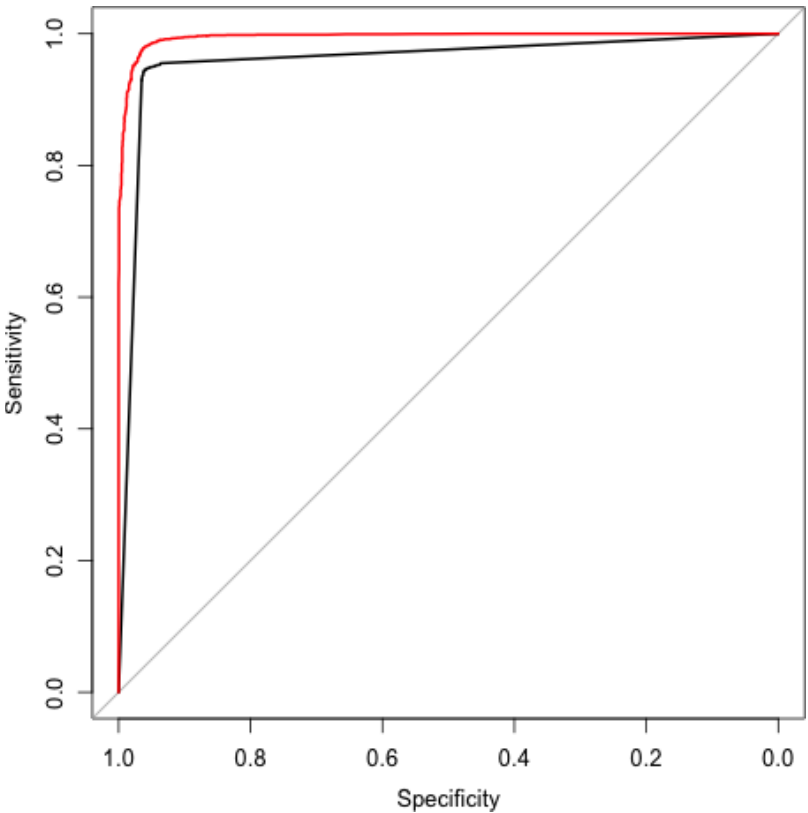


Fig. 10. Receiver operating characteristic curves: —, Gaussian model; —, NIG model

the gradient of log-transformed GFR. This comparison is summarized by using the receiver operating characteristic curves in Fig. 10. Each curve is constructed by varying the predictive probability threshold at which referral is triggered and calculating the corresponding proportions of correct and incorrect (according to the clinical guideline) referrals and non-referrals. There is a substantial increase in predictive power, according to all measures of accuracy, when the correct model, the NIG model, is used.

8. Software

We have implemented the methodology that is presented in this paper in the R package `ngme`. A

development version of the package is available from <https://bitbucket.org/davidbolin/ngme>. The package includes functions for parameter estimation and for subject level prediction using the class of models defined by expression (9), with the following features.

- (a) Any linear model can be specified for the regression term $\mathbf{x}_{ij}^T \boldsymbol{\beta}$ and for the subject level random effect $\mathbf{d}_{ij}^T \mathbf{U}_i$, using the standard R model formula syntax.
- (b) The random-effects distribution can be chosen as normal or NIG.
- (c) The covariance structure of the $W_i(t)$ can be specified as a stationary, exponentially correlated process (the fully asymmetric version), as a symmetric or asymmetric Matérn model with smoothness 1.5, or as a non-stationary integrated random walk or omitted altogether to give non-Gaussian versions of the Laird–Ware model. The distribution for the process can be specified as normal, NIG, GAL or Cauchy.
- (d) The distribution of the measurement error terms can be specified as normal, NIG or t .
- (e) Subject level predictions can be obtained either through nowcasting (conditioning on a subject's past and current measurement data), smoothing (conditioning on all of a subject's data) or forecasting (conditioning on all of a subject's past data).

The generic R functions, `print`, `summary`, `plot`, `fitted` and `residuals` are available for the estimation and prediction functions, and the renal data set is included. We plan to extend the package's functionality to a wider range of models for the stochastic process component $W_i(t)$, including a general Matérn correlation structure. The package also has support for estimation of non-Gaussian models for spatial data as we discuss further in the next section.

9. Discussion

The Gaussian version of the linear mixed model (9) represents the standard approach to analysing real-valued repeated measurement data. Typically, the simplified version without the Gaussian process term $W_i(t)$ suffices when the number of follow-up times per subjects is small, whereas the version with the $W_i(t)$ term, often in conjunction with a simple random intercept U_i in place of the general term $\mathbf{d}_{ij}^T \mathbf{U}_i$, usually gives a better fit to data with long follow-up sequences. Concerns have often been raised about the legitimacy of the Gaussian assumption, and in particular about the consequences of fitting Gaussian models when elements of the underlying process have longer-than-Gaussian tails or skewness. This has led to an extensive literature, which we reviewed in Section 2. However, to the best of our knowledge the current paper is the first to provide a flexible implementation in which departure from Gaussianity can be assessed independently for each of the three stochastic components of model (9).

In our reanalysis of the cystic fibrosis data, inferences on fixed effects showed only small changes when non-Gaussian behaviour is taken into account. Our reanalysis of the renal data also finds evidence of non-Gaussian behaviour, which in this case matters more, because it has a material effect on predictive inference, and hence on the point at which an individual patient in primary care would be considered for referral to secondary care.

We have emphasized the importance of building a computationally efficient algorithm for routine maximization of the likelihood. This is especially useful for data sets containing many subjects. Arguably, computational efficiency is of secondary importance in confirmatory analysis. Once the statistical analysis protocol has been determined, it matters little whether it takes minutes, hours or days of computing time to analyse a data set that typically will have taken weeks, months or years to collect. However, during the iterative model building cycle that characterizes exploratory data analysis, an inability to fit and compare different models in realtime is a severe impediment.

The applications that were described in Section 6 show that the subsampling scheme that was introduced in Section 4.4 can perform very well. One topic of future research is a more thorough investigation of how to optimize the subsampling.

Generalized linear mixed models provide a framework for handling non-Gaussian sampling distributions. This form of non-Gaussian behaviour is complementary to the kind of non-Gaussian process behaviour that we have addressed in this paper. A natural extension to our proposed models would be to generalized linear mixed models for binary or count data with non-Gaussian random effects. However, non-Gaussian behaviour will naturally be more difficult to detect from count or binary data than from measurement data. Binary data in particular can be considered as a heavily censored version of measurement data. For example, a logistic regression model can be interpreted as a linear regression model for a real-valued response Y in which only the sign of Y is observed.

Clinical repeated measurement data are often coupled with time-to-event outcomes, e.g. death. So-called *joint models* for repeated measurement and time-to-event outcomes have been widely studied; for a recent book length account, see Rizopoulos (2012). However, essentially all of this literature assumes that any random-effect components are Gaussian. A natural way of extending the methodology that is presented in this paper to joint modelling problems, by analogy with much of the current literature on Gaussian joint models, would be to combine the linear mixed model (9) with a log-linear Cox process model for the time-to-event outcome, in which the stochastic process $W_i(t)$ in the repeated measurement submodel is correlated with a second stochastic process, $W_i^*(t)$ say, such that $\exp\{W_i^*(t)\}$ constitutes a time-dependent frailty for the i th subject.

Another possible extension of the methodology that is presented in this paper would be to multivariate settings, in which more than one repeated biomarker measurement is collected for each patient, sometimes with different follow-up schedules for different biomarkers. Such models could be constructed similarly to the multivariate random fields in Bolin and Wallin (2020). The models of Bolin and Wallin (2020) are not considered in a longitudinal setting but can naturally be extended to this case. This could then be viewed as an extension of the models that are considered in this work, where the temporal stochastic process is replaced by a random field. The `ngme` package has support for such spatial and multivariate models, both for the longitudinal setting and for the classical geostatistical setting.

Acknowledgements

This work has been partially supported by the Swedish Research Council under grant 201604187. The authors are grateful for the valuable comments from reviewers and editors, which resulted in a much improved manuscript. We also thank Kristin Kirchner for helpful discussions regarding the Petrov–Galerkin method.

Appendix A: Details on the parameter estimation and sampling

A.1. Gibbs sampling

In this section we derive the two conditional distributions that are required for the Gibbs sampler. The first of these is the distribution of $\mathbf{X}_i = (\mathbf{U}_i, \mathbf{W}_i)$ given \mathbf{Y}_i and $\mathbf{V}_i = (\mathbf{V}_i^Z, \mathbf{V}_i^U, \mathbf{V}_i^W)$, and the second is the distribution of \mathbf{V}_i given \mathbf{X}_i and \mathbf{Y}_i . Using the specification of the hierarchical model from Section 4.1 we have that $\mathbf{X}_i | \mathbf{V}_i \sim N(\mathbf{b}_i, \mathbf{Q}_i^{-1})$, where

$$\mathbf{b}_i = \begin{pmatrix} -\boldsymbol{\mu}^U + \boldsymbol{\mu}^U \mathbf{V}_i^U \\ \mathbf{K}^{-1}(-1 + \mathbf{V}_i^W) \boldsymbol{\mu}^W \end{pmatrix}.$$

$$\mathbf{Q}_i = \begin{pmatrix} \frac{1}{V_i^U} \Sigma^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{K}^T \text{diag}\left(\frac{1}{V_i^W}\right) \mathbf{K} \end{pmatrix}.$$

We introduce $\tilde{\mathbf{Y}}_i = \mathbf{Y}_i - \mathbf{x}_i^T \boldsymbol{\beta} + \mu^Z - \mu^Z \mathbf{V}_i^Z$ and $\mathbf{G}_i = (\mathbf{d}_i^T, \mathbf{A}_i^T)^T$. We then have $\tilde{\mathbf{Y}}_i | \mathbf{X}_i, \mathbf{V}_i \sim N\{\mathbf{G}_i \mathbf{X}_i, \sigma^2 \text{diag}(\mathbf{V}_i^Z)\}$ and straightforward calculations using properties of the multivariate normal distributions give that $\mathbf{X}_i | \mathbf{Y}_i, \mathbf{V}_i \sim N(\tilde{\mathbf{b}}_i, \tilde{\mathbf{Q}}_i)$, where

$$\tilde{\mathbf{b}}_i = \tilde{\mathbf{Q}}_i^{-1} \left(\mathbf{Q}_i^{-1} \mathbf{b}_i + \sigma^{-2} \mathbf{G}_i^T \text{diag}\left(\frac{1}{V_i^Z}\right) \tilde{\mathbf{Y}}_i \right), \quad \tilde{\mathbf{Q}}_i = \mathbf{Q}_i + \sigma^{-2} \mathbf{G}_i^T \text{diag}\left(\frac{1}{V_i^Z}\right) \mathbf{G}_i.$$

To compute the distribution of \mathbf{V}_i given \mathbf{X}_i and \mathbf{Y}_i , we use the following proposition regarding the convolution between a GIG variable and a Gaussian variable.

Proposition 1. Let $V \sim \text{GIG}(p, a, b)$ and $\mathbf{Y} | V \sim N(\boldsymbol{\beta} + \mu \mathbf{v}, v \Sigma)$ where $\mathbf{Y} \in \mathbb{R}^n$. Then

$$V | \mathbf{Y} \sim \text{GIG}\{v; p - 0.5n, a + \mu^T \Sigma^{-1} \mu, b + (\mathbf{Y} - \boldsymbol{\beta})^T \Sigma^{-1} (\mathbf{Y} - \boldsymbol{\beta})\}.$$

The proof is straightforward and hence has been omitted. Now, note that the density of $\mathbf{V}_i | \mathbf{X}_i, \mathbf{Y}_i$ factorizes as $f(\mathbf{V}_i | \mathbf{X}_i, \mathbf{Y}_i) = f(\mathbf{V}_i^Z | \mathbf{X}_i, \mathbf{Y}_i) f(\mathbf{V}_i^U | \mathbf{U}_i) f(\mathbf{V}_i^W | \mathbf{W}_i)$. If $V_i^U \sim \text{GIG}(p^U, a^U, b^U)$, $V_{ij}^Z \sim \text{GIG}(p^Z, a^Z, b^Z)$, and $V_{ij}^W \sim \text{GIG}(p^W, a^W, b^W)$, then the proposition gives that the three independent distributions are

$$\begin{aligned} V_i^U | \mathbf{U}_i &\sim \text{GIG} \left\{ p^U - \frac{d^U}{2}, a^U + (\mu^U)^T \Sigma^{-1} \mu^U, b^U + (\mathbf{U}_i + \mu^U)^T \Sigma^{-1} (\mathbf{U}_i + \mu^U) \right\}, \\ \mathbf{V}_{ij}^Z | \mathbf{X}_i, \mathbf{Y}_i &\sim \text{GIG} \left\{ p^Z - 0.5, a^Z + \left(\frac{\mu^Z}{\sigma} \right)^2, b^Z + \frac{(Y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{d}_{ij}^T \mathbf{U}_i - \mathbf{A}_{ij}^T \mathbf{W}_i)^2}{\sigma^2} \right\}, \\ \mathbf{V}_{ij}^W | \mathbf{W}_i &\sim \text{GIG} \{ p^W - 0.5, a^W + (\mu^W)^2, b^W + (\mathbf{K} \mathbf{W}_i + \mathbf{h} \mu^W)^2 \}. \end{aligned}$$

A.2. The gradient of the likelihood

In this section we derive the gradient that is needed for estimating the parameters in the model from Section 4.1. We shall use the notation from Appendix A.1 and also assume that $E[V_i^U] = 1$, $E[V_i^Z] = \mathbf{1}$ and $E[V_i^Z] = \mathbf{h}_i$, which was previously assumed for parameter identifiability. Recall that the goal is to evaluate the gradient in equation (18), which can be written as a sum over $\nabla_{\boldsymbol{\Theta}} L_i(\boldsymbol{\Theta}; \mathbf{Y}_i)$, which we approximate by using MC integration as

$$\nabla_{\boldsymbol{\Theta}} L_i(\boldsymbol{\Theta}; \mathbf{Y}_i) \approx \frac{1}{N_{\text{MC}}} \sum_{j=1}^{N_{\text{MC}}} \nabla_{\boldsymbol{\Theta}} L_i(\boldsymbol{\Theta}; \mathbf{Y}_i, \mathbf{V}_i^{(j)})$$

where $\mathbf{V}_i^{(j)}$ are samples from $\mathbf{V}_i | \mathbf{Y}_i$ obtained by using the Gibbs sampler above. Thus, what we must derive here is $\nabla_{\boldsymbol{\Theta}} L_i(\boldsymbol{\Theta}; \mathbf{Y}_i, \mathbf{V}_i^{(j)})$. To do this, we shall use that $\nabla_{\boldsymbol{\Theta}} L_i(\boldsymbol{\Theta}; \mathbf{Y}_i, \mathbf{V}_i^{(j)}) = E_{\mathbf{X}_i}[\nabla_{\boldsymbol{\Theta}} L_i(\boldsymbol{\Theta}; \mathbf{Y}_i, \mathbf{X}_i, \mathbf{V}_i^{(j)}) | \mathbf{Y}_i, \mathbf{V}_i]$.

To simplify the notation, we let $L_i(\tilde{\boldsymbol{\Theta}}; \mathbf{X}_i, \mathbf{V}_i, \mathbf{Y}_i)$ denote the complete log-likelihood for the i th patient seen as a function of the parameters $\boldsymbol{\Theta}$, and similarly let $L_i(\boldsymbol{\Theta} | \mathbf{V}_i, \mathbf{Y}_i)$ denote the log-likelihood conditioned on the variance components. To derive the gradients, we also need the following notation from matrix calculus. The *vec* operator transforms a matrix into a vector by stacking its columns. The *vech* operator also transforms an $n \times n$ matrix into a vector but removes all the subdiagonal elements. Finally, the duplication matrix \mathbf{D}_n is such that, for any symmetric matrix \mathbf{A} , $\mathbf{D}_n \text{vech}(\mathbf{A}) = \text{vec}(\mathbf{A})$.

We start by deriving the gradients for the fixed effect and the asymmetric parameters. Let $(\boldsymbol{\beta}, \boldsymbol{\mu}) = (\boldsymbol{\beta}, \mu^U, \mu^Z, \mu^W)$ and $\mathbf{B}_i = (\mathbf{x}_i, \mathbf{d}_i(-1 + V_i^U), (-1 + V_i^Z), \mathbf{A}_i \mathbf{K}^{-1}(-\mathbf{h}_i + \mathbf{V}_i^W))$. Using the model definition from Section 4.1, we have that

$$L_i\{(\boldsymbol{\beta}, \boldsymbol{\mu}), \sigma; \mathbf{X}_i, \mathbf{V}_i, \mathbf{Y}_i\} = -\frac{1}{2\sigma_z^2} (\mathbf{y}_i - \mathbf{B}_i(\boldsymbol{\beta}, \boldsymbol{\mu}) - \mathbf{G}_i \tilde{\mathbf{X}}_i)^T \text{diag}\left(\frac{1}{V_i^Z}\right) (\mathbf{y}_i - \mathbf{B}_i(\boldsymbol{\beta}, \boldsymbol{\mu}) - \mathbf{G}_i \tilde{\mathbf{X}}_i),$$

where $\tilde{\mathbf{X}}_i = \mathbf{X}_i - \mathbf{b}_i$. Thus the gradient with respect to (β, μ) equals

$$\nabla_{[\beta, \mu]} L_i(\Theta; \mathbf{Y}_i, \mathbf{V}_i, \mathbf{X}_i) = \frac{1}{\sigma_z^2} \mathbf{B}_i^T \text{diag}\left(\frac{1}{\mathbf{V}_i^Z}\right) (\mathbf{y}_i - \mathbf{B}_i(\beta, \mu) - \mathbf{G}_i \tilde{\mathbf{X}}_i).$$

Using that the expected value of \mathbf{X}_i given \mathbf{V}_i and \mathbf{Y}_i is $\tilde{\mathbf{b}}_i$ (see Appendix A.1) we obtain that

$$\nabla_{(\beta, \mu)} L_i(\Theta; \mathbf{Y}_i, \mathbf{V}_i) = \frac{1}{\sigma_z^2} \mathbf{B}_i^T \text{diag}\left(\frac{1}{\mathbf{V}_i^Z}\right) (\mathbf{y}_i - \mathbf{B}_i(\beta, \mu) - \mathbf{G}_i(\tilde{\mathbf{b}}_i - \mathbf{b}_i)).$$

The gradient for noise variance σ^2 , of the complete log-likelihood, equals

$$\nabla_{\sigma} \log\{L_i(\Theta; \mathbf{y}_i, \mathbf{X}_i, \mathbf{V}_i)\} = -\frac{m_i}{\sigma} + \frac{1}{\sigma^3} \left(\mathbf{e}_i \frac{1}{\mathbf{V}_i^Z} \right)^T \mathbf{e}_i,$$

where m_i is the number observations of patient i and $\mathbf{e}_i = \mathbf{y}_i - \mathbf{B}_i(\beta, \mu) - \mathbf{G}_i \tilde{\mathbf{X}}_i$. We could compute $\nabla_{\sigma} \log\{L_i(\Theta; \mathbf{y}_i, \mathbf{V}_i)\}$ by taking the expectation with respect to \mathbf{X}_i , but in our implementation we simply use the values of \mathbf{X}_i from the Gibbs sampler to approximate this expected value by using MC integration, the reason being that the estimation of σ is so simple compared with the other parameters that it was not worth the additional effort to implement the analytical gradient for this parameter.

To derive the gradient with respect to the covariance matrix of the random effects, we first note that

$$\begin{aligned} L_i(\Sigma; \mathbf{Y}_i, \mathbf{X}_i, \mathbf{V}_i) &= -\frac{1}{2} \log(|\Sigma|) - \frac{1}{2V_i^U} (\mathbf{U}_i^U + (1 - V_i^U)\mu^U)^T \Sigma^{-1} (\mathbf{U}_i^U + (1 - V_i^U)\mu^U) \\ &= \frac{1}{2V_i^U} (\tilde{\mathbf{U}}_i^U)^T \Sigma^{-1} \tilde{\mathbf{U}}_i^U, \end{aligned}$$

where $\tilde{\mathbf{U}}_i^U = \mathbf{U}_i^U + (1 - V_i^U)\mu^U$. Denoting $\mathbf{M}_i = \tilde{\mathbf{U}}_i^U (\tilde{\mathbf{U}}_i^U)^T$, \mathbf{K}_{dd} the commutation matrix and \mathbf{D}_d the duplication matrix (Magnus and Neudecker (2007), pages 389–390), the gradient for the variance matrix Σ is

$$\nabla_{\text{vech}(\Sigma)} L_i(\Theta; \mathbf{Y}_i, \mathbf{X}_i, \mathbf{V}_i) = \frac{1}{2} \mathbf{D}_d^T (\Sigma^{-1} \otimes \Sigma^{-1}) \text{vec}\left(\frac{\mathbf{M}_i}{V_i^U} - \Sigma\right).$$

Defining $\tilde{\Sigma} = \tilde{\mathbf{Q}}_i^{-1}$ and $\tilde{\mathbf{b}} = \tilde{\mathbf{b}}_i$, the expected value equals

$$\nabla_{\text{vech}(\Sigma)} L_i(\Theta; \mathbf{Y}_i, \mathbf{V}_i) = \frac{1}{2} \mathbf{D}_d^T (\Sigma^{-1} \otimes \Sigma^{-1}) \text{vec}\left(\frac{\tilde{\Sigma}_{1:d, 1:d} + \tilde{\mathbf{b}}_{1:d} \tilde{\mathbf{b}}_{1:d}^T}{V_i^U} - \Sigma\right).$$

For a generic parameter θ of the differential operator that is used to define the process, we have that

$$L_i(\theta; \mathbf{Y}_i, \mathbf{X}_i, \mathbf{V}_i) = \log(|\mathbf{K}|) - \frac{1}{2} (\mathbf{K} \mathbf{W}_i + (\mathbf{h} - \mathbf{V}_i^W) \mu^W)^T \text{diag}(\mathbf{V}_i^W)^{-1} (\mathbf{K} \mathbf{W}_i + (\mathbf{h} - \mathbf{V}_i^W) \mu^W).$$

Thus, defining $(\mathbf{K}\theta)_{ij} = dK_{ij}/d\theta$, the gradient equals

$$\nabla_{\theta} \log\{L_i(\Theta)\} = \text{tr}(\mathbf{K}\theta \mathbf{K}^{-1}) - \mathbf{W}_i^T \mathbf{K}_{\theta}^T \text{diag}(\mathbf{V}_i^W)^{-1} (\mathbf{K} \mathbf{W}_i + (\mathbf{h} - \mathbf{V}_i^W) \mu^W).$$

What remains is to compute the gradient with respect to the variance mixing parameters. For these parameters the complete log-likelihood is entirely determined by the specified distribution of the variance mixing variables. Thus there is no generic form; instead we present the three main distributions that we have considered and their resulting gradients. The three distributions are the gamma distribution with density $f(v; p, b) = \Gamma(p)^{-1} a^p v^{p-1} \exp(-av)$, the inverse Gaussian distribution with density $f(v; a, b) = b^{1/2} (2\pi)^{-1/2} v^{-3/2} \exp\{-av/2 - b/(2v) + \sqrt{(ab)}\}$, and the inverse gamma distribution with density $f(v; p, a) = \Gamma(p)^{-1} a^p v^{-p-1} \exp(-b/v)$. The resulting gradients for the noise parameters are given in Table 8, the random-effect parameters in Table 9 and the processes parameters in Table 10. In Tables 8–10, ψ is the digamma function and ψ_1 is the trigamma function. Note that we use a non-standard form of the t -distribution where ν is half of the degrees of freedom and the parameterization is chosen so that a sym-

Table 8. Distribution, gradient and observed Fisher information for the mixing variable of the noise

Distribution	Distribution of V^Z	$\nabla L_i(\cdot; \mathbf{Y}_i, \mathbf{V}_i)$	$E[\nabla^2 L_i(\cdot; \mathbf{Y}_i, \mathbf{V}_i)]$
t	$\text{IGam}(\nu, \nu - 1)$	$m_i \left\{ \log(\nu - 1) + \frac{\nu}{\nu - 1} - \psi(\nu) \right\} - \sum_{j=1}^{m_i} \log(V_{ij}^Z) - \frac{1}{V_{ij}^Z}$	$m_i \left\{ \frac{1}{\nu - 1} - \psi_1(\nu) \right\}$
NIG	$\text{IG}(\nu, \nu)$	$-\frac{1}{2} \left(m_i \nu^{-1} - \sum_{j=1}^{m_i} \frac{1}{V_{ij}^Z} + V_{ij}^Z - 1 \right)$	$\frac{m_i}{2\nu^2}$
GAL	$\text{Gam}(\lambda, \lambda)$	$m_i \log(\lambda) + m_i - m_i \psi(\lambda) + \sum_{j=1}^{m_i} \log(V_{ij}^Z) - V_{ij}^Z$	$m_i \left\{ \frac{1}{\lambda} - \psi_1(\lambda) \right\}$

Table 9. Distribution, gradient and observed Fisher information for the mixing variable of the random effect

Distribution	Distribution of V^U	$\nabla L_i(\cdot; \mathbf{Y}_i, \mathbf{V}_i)$	$E[\nabla^2 L_i(\cdot; \mathbf{Y}_i, \mathbf{V}_i)]$
t	$\text{IGam}(\nu, \nu - 1)$	$\log(\nu - 1) + \frac{\nu}{\nu - 1} - \psi(\nu) - \log(V_i^U) - \frac{1}{V_i^U}$	$\frac{1}{\nu - 1} - \psi_1(\nu)$
NIG	$\text{IG}(\nu, \nu)$	$-\frac{1}{2} \left(\nu^{-1} - \frac{1}{V_i^U} + V_i^U - 1 \right)$	$\frac{1}{2\nu^2}$
GAL	$\text{Gam}(\lambda, \lambda)$	$\log(\lambda) + -\psi(\lambda) + \log(V_i^U) - V_i^Z$	$\frac{1}{\lambda} - \psi_1(\lambda)$

Table 10. Distribution, gradient and observed Fisher information for the mixing variable for the processes†

Distribution	Distribution of V^Z	$\nabla L_i(\cdot; \mathbf{Y}_i, \mathbf{V}_i)$	$E[\nabla^2 L_i(\cdot; \mathbf{Y}_i, \mathbf{V}_i)]$
NIG	$\text{IG}(\nu, \nu h_{ij}^2)$	$-\sum_{j=1}^{n_i} \frac{1}{2} \left(\nu^{-1} - \frac{h_{ij}^2}{V_{ij}^W} + V_{ij}^W - h_{ij} \right)$	$\frac{n_i}{2\nu^2}$
GAL	$\text{Gam}(h_{ij}\lambda, \lambda)$	$\sum_{j=1}^{n_i} h_{ij} \{ \log(\lambda) + 1 \} - \psi(h_{ij}\lambda) + h_{ij} \log(V_{ij}^U) - V_{ij}^U$	$\sum_{j=1}^{n_i} h_{ij} \left\{ \frac{1}{\lambda} - \psi_1(h_{ij}\lambda) \right\}$

†Here h_{ij} is the integral of the basis function (see Appendix B).

metric version has variance 1. For $\nu < 2$ this is not possible (since the variance is unbounded) and one can then instead use the parameterization $\text{IGam}(\nu/2, \nu/2 + 1)$ which puts the mode of the IGam density at 1.

A.2.1. Joint Fisher information for mixed effects parameters

When computing Wald-type confidence intervals based on the Fisher information matrix, one should ideally compute it from the joint Hessian for all parameters. A simpler alternative which we make is to compute the joint Fisher information matrix only for the mixed effect parameters, which means that we do not take the uncertainty of the other parameters into account. This simplifies the implementation greatly and should in most scenarios have little effect since the other parameters converge much faster than the mixed effect parameters. The reason for this is that the random effects vary between individuals, so one individual can be seen as an observation whereas the other parameters receive information also from all longitudinal observations for each patient.

Let $\Theta_m = (\beta, \mu, \text{vech}(\Sigma), \nu)$ be the vector of all parameters for a model with NIG-distributed mixed effects and let $L_i(\Theta_m)$ denote the complete likelihood $L_i(\Theta_m; \mathbf{Y}_i, \mathbf{V}_i, \mathbf{X}_i)$. The negative Hessian of the likelihood for patient i (the observed Fisher information is the Hessian at the mode) for these parameters can be expressed as

$$\begin{aligned}
 -\mathbf{H}_i = & -E_{\mathbf{V}_i}[E_{\mathbf{X}_i}[\nabla_{\Theta_m}^2 L_i(\Theta_m)|\mathbf{Y}_i, \mathbf{V}_i]|\mathbf{Y}_i] - E_{\mathbf{V}_i}[E_{\mathbf{X}_i}[\nabla_{\Theta_m} L_i(\Theta_m)\nabla_{\Theta_m} L_i(\Theta_m)^T|\mathbf{Y}_i, \mathbf{V}_i]|\mathbf{Y}_i] \\
 & + E_{\mathbf{V}_i}[E_{\mathbf{X}_i}[\nabla_{\Theta_m} L_i(\Theta_m)|\mathbf{Y}_i, \mathbf{V}_i]|\mathbf{Y}_i]E_{\mathbf{V}_i}[E_{\mathbf{X}_i}[\nabla_{\Theta_m} L_i(\Theta_m)|\mathbf{Y}_i, \mathbf{V}_i]|\mathbf{Y}_i]^T.
 \end{aligned}$$

We shall now derive the inner expectations with respect to \mathbf{X} (the outer are computed through MCMC sampling). We shall utilize the distribution derived for the Gibbs sampling in Appendix A.1, but for simplicity drop the index of the mean and covariance: $\tilde{\mathbf{X}} \sim N(\mathbf{b} + \mathbf{b}, \mathbf{Q}^{-1}) = N(\mathbf{b}^*, \Sigma)$; going further we shall always work with the i th patient and therefore drop the index in the notation.

For the derivations, we shall need the following lemma.

Lemma 1. Let

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_A \\ \mathbf{X}_B \end{pmatrix} \sim N\left\{ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_A \\ \boldsymbol{\mu}_B \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BB} & \Sigma_{BB} \end{pmatrix} \right\}$$

be a d -dimensional normal distribution where A and B are subsets of $\{1, 2, \dots, d\}$ such that $A \cup B = \{1, 2, \dots, d\}$. Then

$$\begin{aligned} E[\text{vec}(\mathbf{X}_A \mathbf{X}_A^T) \mathbf{X}_B^T] &= E[\mathbf{X}_A \otimes \mathbf{X}_A \otimes \mathbf{X}_B^T] \\ &= \boldsymbol{\mu}_B^T \otimes \boldsymbol{\mu}_A \otimes \boldsymbol{\mu}_A + \boldsymbol{\mu}_A \otimes \Sigma_{AB} + \Sigma_{AB} \otimes \boldsymbol{\mu}_A + \text{vec}(\Sigma_{A,A}) \boldsymbol{\mu}_B^T. \end{aligned} \quad (24)$$

Proof. From corollary 2.2.7.2 in Kollo and von Rosen (2006) it follows that

$$E[\mathbf{X} \otimes \mathbf{X} \otimes \mathbf{X}^T] = \boldsymbol{\mu}^T \otimes \boldsymbol{\mu} \otimes \boldsymbol{\mu} + \boldsymbol{\mu} \otimes \Sigma + \Sigma \otimes \boldsymbol{\mu} + \text{vec}(\Sigma) \boldsymbol{\mu}^T. \quad (25)$$

To link this result with equation (24) we define the two matrices, \mathbf{B}_A and \mathbf{B}_B , such that $\mathbf{X}_A = \mathbf{B}_A \mathbf{X}$, and $\mathbf{X}_B = \mathbf{B}_B \mathbf{X}$. Now it follows that $\text{vec}(\mathbf{X}_A \mathbf{X}_A^T) = \text{vec}(\mathbf{B}_A \mathbf{X} \mathbf{X}^T \mathbf{B}_A^T) = \mathbf{B}_A \otimes \mathbf{B}_A \text{vec}(\mathbf{X} \mathbf{X}^T)$, and the left-hand side of equation (24) can be rewritten as

$$E[\text{vec}(\mathbf{X}_A \mathbf{X}_A^T) \mathbf{X}_B^T] = \mathbf{B}_A \otimes \mathbf{B}_A E[\mathbf{X} \otimes \mathbf{X} \otimes \mathbf{X}^T] \mathbf{B}_B^T.$$

Joining the equation above with equation (25) gives the following three terms from which we shall extract the final result:

$$\begin{aligned} (\mathbf{B}_A \otimes \mathbf{B}_A)(\boldsymbol{\mu}^T \otimes \boldsymbol{\mu} \otimes \boldsymbol{\mu}) \mathbf{B}_B^T &= (\mathbf{B}_B \boldsymbol{\mu})^T \otimes \mathbf{B}_A \boldsymbol{\mu} \otimes \mathbf{B}_A \boldsymbol{\mu} = \boldsymbol{\mu}_B^T \otimes \boldsymbol{\mu}_A \otimes \boldsymbol{\mu}_A, \\ (\mathbf{B}_A \otimes \mathbf{B}_A)(\boldsymbol{\mu} \otimes \Sigma) \mathbf{B}_B^T &= \mathbf{B}_A \boldsymbol{\mu} \otimes \mathbf{B}_A \Sigma \mathbf{B}_B^T = \boldsymbol{\mu}_A \otimes \Sigma_{AB}, \\ (\mathbf{B}_A \otimes \mathbf{B}_A)(\text{vec}(\Sigma) \boldsymbol{\mu}^T) \mathbf{B}_B^T &= \text{vec}(\mathbf{B}_A \Sigma \mathbf{B}_A^T) (\mathbf{B}_B \boldsymbol{\mu})^T = \text{vec}(\Sigma_{A,A}) \boldsymbol{\mu}_B^T. \end{aligned} \quad \square$$

Several of the required gradients are straightforward to derive and we therefor omit most details. For example, some simple second derivatives that are needed are

$$\begin{aligned} \nabla_{[\beta, \mu]}^2 L_i(\boldsymbol{\Theta}_m) &= -\frac{1}{\sigma^2} \frac{1}{V^Z} \mathbf{B}_i^T \mathbf{B}_i, \\ \nabla_{\text{vech}(\Sigma)}^2 L_i(\boldsymbol{\Theta}_m) &= -\frac{1}{2} \mathbf{D}_d^T \mathbf{K}_{dd} \left(\Sigma^{-1} \otimes \Sigma^{-1} \left(\frac{\mathbf{M}_i}{V_i^U} - \Sigma \right) \Sigma^{-1} + \Sigma^{-1} \left(\frac{\mathbf{M}_i}{V_i^U} \right) \Sigma^{-1} \otimes \Sigma^{-1} \right) \mathbf{D}_d, \\ E[\nabla_{\text{vech}(\Sigma)}^2 L_i(\boldsymbol{\Theta}_m)] &= -\frac{1}{2} \mathbf{D}_d^T \mathbf{K}_{dd} (\Sigma^{-1} \otimes \Sigma^{-1}) \mathbf{D}_d. \end{aligned}$$

Likewise, for the parameter of the mixing component ν , the gradient does not depend on $\mathbf{X}|\mathbf{V}$ and its expected value is hence easily evaluated from the gradient. See Tables 8, 9 and 10.

We now present the more difficult components to compute. We start with evaluating the expected value of the outer product of $\nabla_{\text{vech}(\Sigma)} L_i(\boldsymbol{\Theta}_m)$ which we split into several parts. The gradient is

$$\nabla_{\text{vech}(\Sigma)} L_i(\boldsymbol{\Theta}_m) = \mathbf{H}_\Sigma \text{vec}\left(\frac{\mathbf{M}}{V^U}\right) - \mathbf{H}_\Sigma \text{vec}(\Sigma),$$

and thus the outer product is

$$\nabla_{\text{vech}(\Sigma)} L_i(\boldsymbol{\Theta}_m) \nabla_{\text{vech}(\Sigma)} L_i(\boldsymbol{\Theta}_m)^T = \mathbf{H}_\Sigma (\mathbf{K}_1 - \mathbf{K}_2 - \mathbf{K}_2^T + \mathbf{K}_3) \mathbf{H}_\Sigma^T,$$

where $\mathbf{K}_1 = \text{vec}(\mathbf{M}_i / V_i^U) \text{vec}(\mathbf{M} / V^U)^T$, $\mathbf{K}_2 = \text{vec}(\mathbf{M} / V^U) \text{vec}(\Sigma)^T$ and $\mathbf{K}_3 = \text{vec}(\Sigma) \text{vec}(\Sigma)^T$. To compute the expectation of the outer product we need the expectation of $\text{vec}(\mathbf{M})$ and $\text{vec}(\mathbf{M}) \text{vec}(\mathbf{M})^T$. First

$$E_{\mathbf{X}} \left[\frac{1}{V^U} \text{vec}(\mathbf{M}) | \mathbf{Y}, \mathbf{V} \right] = \frac{1}{V^U} \text{vec}(E_{\mathbf{X}}[\tilde{\mathbf{X}}_{1:d} \tilde{\mathbf{X}}_{1:d}^T]) = \frac{1}{V^U} \text{vec}\{\tilde{\Sigma}_{1:d, 1:d} + \tilde{\mathbf{b}}_{1:d}^* (\tilde{\mathbf{b}}_{1:d}^*)^T\}.$$

To derive the expectation of \mathbf{K}_1 we use theorem 4.3 of Magnus and Neudecker (1979):

$$\begin{aligned} E_{\mathbf{X}} \left[\frac{1}{(V^U)^2} \text{vec}(\mathbf{M}) \text{vec}(\mathbf{M})^T | \mathbf{Y}, \mathbf{V} \right] &= \frac{1}{(V^U)^2} E_{\mathbf{X}}[\tilde{\mathbf{X}}_{1:d} \tilde{\mathbf{X}}_{1:d}^T \otimes \tilde{\mathbf{X}}_{1:d} \tilde{\mathbf{X}}_{1:d}^T | \mathbf{Y}, \mathbf{V}] \\ &= \frac{1}{(V^U)^2} (\mathbf{I} + \mathbf{K}_d) (\tilde{\Sigma}_{1:d, 1:d} \otimes \tilde{\Sigma}_{1:d, 1:d} + \tilde{\Sigma}_{1:d, 1:d} \otimes \tilde{\mathbf{b}}_{1:d}^* (\tilde{\mathbf{b}}_{1:d}^*)^T \\ &\quad + \tilde{\mathbf{b}}_{1:d}^* (\tilde{\mathbf{b}}_{1:d}^*)^T \otimes \tilde{\Sigma}_{1:d, 1:d}) + E_{\mathbf{X}} \left[\frac{1}{V^U} \text{vec}(\mathbf{M}) | \mathbf{Y}, \mathbf{V} \right] E_{\mathbf{X}} \left[\frac{1}{V^U} \text{vec}(\mathbf{M}) | \mathbf{Y}, \mathbf{V} \right]^T, \end{aligned}$$

where \mathbf{K}_{dd} again denotes the commutation matrix.

We continue with the outer product of $\nabla_{[\beta, \mu]} L_i(\Theta_m)$:

$$\nabla_{[\beta, \mu]} L_i(\Theta_m) \nabla_{[\beta, \mu]} L_i(\Theta_m)^T = \mathbf{a} \mathbf{a}^T - \mathbf{a} \mathbf{A}^T \mathbf{X}^T - \mathbf{A} \mathbf{X} \mathbf{a}^T + \mathbf{A} \mathbf{X} \mathbf{X}^T \mathbf{A}^T,$$

where

$$\mathbf{a} = \frac{1}{2\sigma_z^2} \mathbf{B}^T \text{diag}\left(\frac{1}{\sqrt{z}}\right) (\mathbf{y} - \mathbf{B}[\beta, \mu])$$

and

$$\mathbf{A} = \frac{1}{2\sigma_z^2} \mathbf{B}^T \text{diag}\left(\frac{1}{\sqrt{z}}\right) \mathbf{G}.$$

The expectations of the terms are easily obtained since $\mathbf{X} \sim N(\tilde{\mathbf{b}}, \tilde{\Sigma})$. To compute the expectation we typically do not compute $\tilde{\Sigma}$ but rather solve linear systems using $\tilde{\mathbf{Q}}$.

The final difficulty is to compute the expectation of the outer product:

$$\begin{aligned} \nabla_{[\beta, \mu]} L_i(\Theta_m) \nabla_{\text{vech}(\Sigma)} L_i(\Theta_m)^T &= \frac{1}{2\sigma_z^2} \mathbf{B}^T \text{diag}\left(\frac{1}{\sqrt{z}}\right) (\mathbf{Y} - \mathbf{B}(\beta, \mu) - \mathbf{G} \tilde{\mathbf{X}}_i) \text{vec}\left(\frac{\mathbf{M}}{V^U} - \Sigma\right)^T \mathbf{H}_{\Sigma}^T \\ &= \tilde{\mathbf{a}} \left(-\text{vec}(\Sigma)^T + \frac{1}{V^U} \text{vec}(\mathbf{M}_i)^T \right) \mathbf{H}_{\Sigma}^T + \mathbf{A} \tilde{\mathbf{X}} \left(\text{vec}(\Sigma)^T - \frac{1}{V^U} \text{vec}(\mathbf{M})^T \right) \mathbf{H}_{\Sigma}^T, \end{aligned}$$

where

$$\tilde{\mathbf{a}} = \mathbf{a} + \frac{1}{2\sigma_z^2} \mathbf{B}^T \text{diag}\left(\frac{1}{\sqrt{z}}\right) \mathbf{G} \mathbf{b}.$$

For most of the terms in this expression we have already calculated the corresponding expectation before. The only new term is $\tilde{\mathbf{X}} \text{vec}(\mathbf{M})^T$, which by lemma 1 is

$$E_{\mathbf{X}}[\tilde{\mathbf{X}} \text{vec}(\mathbf{M})^T | \mathbf{Y}, \mathbf{V}] = \tilde{\mathbf{b}}^* \otimes (\tilde{\mathbf{b}}_{1:d}^*)^T \otimes (\tilde{\mathbf{b}}_{1:d}^*)^T + (\tilde{\mathbf{b}}_{1:d}^*)^T \otimes \tilde{\Sigma}_{\cdot, 1:d} + \tilde{\Sigma}_{\cdot, 1:d} \otimes (\tilde{\mathbf{b}}_{1:d}^*)^T + \tilde{\mathbf{b}}^* \text{vec}(\tilde{\Sigma}_{1:d, 1:d})^T.$$

A.3. Similarity between densities

Finding the parameters for the GH distribution is in general difficult, because the full log-likelihood surface is largely flat, which makes the model parameters almost non-identifiable. A further difficulty is that the boundary of the parameter space often contains, unique, distributions and hence one cannot expect that the parameter will be contained in a compact region of the parameter space. This problem is also true within the subfamilies that were discussed above. For instance, an NIG distribution converges to a Cauchy distribution as $a \rightarrow 0$, and to a Gaussian distribution if $a \rightarrow \infty$ and $b \rightarrow \infty$ at the same rate. Recognizing these limiting cases is important in practice since it can lead to situations where the parameters do not converge.

A remedy for this issue within the subfamilies that we are studying is to fix a compact parameter space and if the parameters converge to the boundary then we re-estimate the parameters in the limiting distribution. We set these boundaries by examining the total variation (TV) distance between pairs of

densities. For illustration, to compare a symmetric NIG distribution with fixed a with a symmetric Cauchy distribution, we calculate

$$\text{TV}_{\text{NIG,CH}}(b_{\text{CH}}, a, b_{\text{NIG}}) = \min_{b_{\text{CH}}} \int |\text{CH}(x; 0, 0, b_{\text{CH}}) - \text{NIG}(x; 0, 0, a, b_{\text{NIG}})| dx.$$

To simplify the calculations that are needed to find the Cauchy distribution $\text{CH}(0, 0, b_{\text{CH}})$ that is closest to the $\text{NIG}(0, 0, a, b_{\text{NIG}})$ distribution we use the following proposition, which shows that it suffices first to find the Cauchy distribution that is closest to $\text{NIG}(0, 0, a, 1)$, and then to rescale the shape parameter by b_{NIG} .

Proposition 2. Let $f_s(x)$ and $g_h(x)$ be two distributions with respect to the Lebesgue measure, with scaling parameters s and h . Then, $\text{TV}(f_s, g_h) = \text{TV}(f_{s/c}, g_{h/c})$ for $c > 0$.

Proof. First note that

$$\text{TV}(f_{s/c}, g_{h/c}) = \frac{1}{2} \int |f_{s/c}(x) - g_{h/c}(x)| dx = \frac{c}{2} \int |f_s(cx) - g_h(cx)| dx.$$

Now use integration by substitution with respect to $\phi(x) = x/c$ to give

$$\frac{c}{2} \int |f_s(cx) - g_h(cx)| dx = \frac{1}{2} \int |f_s(x) - g_h(x)| dx = \text{TV}(f_s, g_h). \quad \square$$

Fig. 11 shows the TV distances between the NIG and Cauchy, and between the NIG and Gaussian distributions, as functions of a . One can now translate the difference between the densities to compare with densities that one is more familiar with. For example, for $a = 0.001$, the TV distance between the NIG and Cauchy distribution is less than that between two Bernoulli distributions whose probabilities differ by 0.002. The same applies to the TV distance between the NIG and normal distributions when $a = 250$.

In `ngme` we set the boundaries of the NIG parameter space to $0.001 \leq a < 250$, and if the parameter space is hit it gives a warning. Of course one needs to be a little cautious with using the limiting distribution,

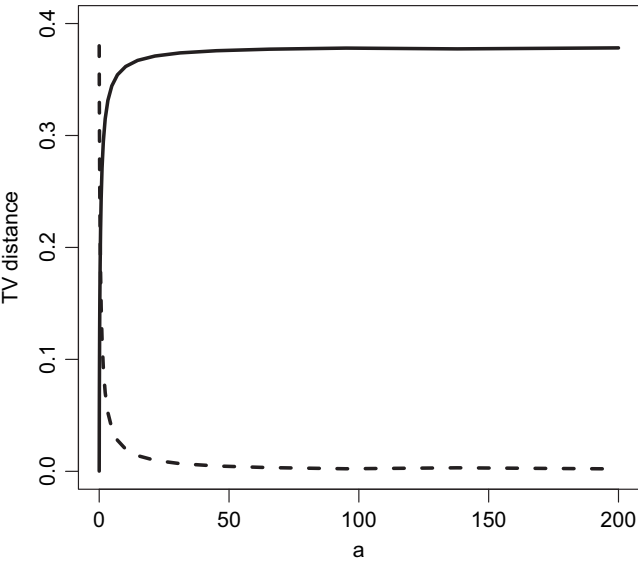


Fig. 11. TV distance between the NIG and Cauchy (— —) and between the NIG and normal (——) distributions for varying a

since, although the TV difference decreases, the tails of the distributions remain different. For instance, the NIG distribution has exponential tails whereas the Cauchy distribution has polynomial tails.

A.4. Pseudocode for the grouped subsampler

Algorithm 1 (Table 11) contains pseudocode describing the group formation of the grouped subsampler.

Appendix B: Discretization

In this section we outline how the stochastic differential equation (13) is discretized. Let $\langle f, g \rangle = \int f(t)g(t)dt$ denote the standard inner product on \mathbb{R} . Recall that we restrict $W(t)$ to a finite interval, $\bar{T} = [-r, t_{\max} + r]$, and impose boundary conditions on the operator to obtain a well-posed problem. The so-called *weak solution* of equation (13) is a function of $W(t)$ that satisfies the equation

$$\langle \psi, \mathcal{D}W \rangle = \langle \psi, dL \rangle, \quad (26)$$

for a specified set of *test functions* $\psi(t)$. Recall that we use the *low rank* approximation (14) and now want to compute the distribution of the weights in this basis expansion. When $\mathcal{D} = \kappa^2 - \partial^2/\partial t^2$, we can use a standard Galerkin finite element discretization; see also Lindgren and Rue (2008). This consists of setting all the test functions to the basis functions, i.e. $\psi_k = \phi_k$ for all k , and computing the W_k by solving the system of equations that is defined by equation (26), i.e. $\mathbf{KW} = \mathbf{L}$, where $L_k = \langle \psi_k, dL \rangle$, and \mathbf{K} is a discretized version of the differential operator \mathcal{D} with elements

$$K_{kk'} = \langle \psi_k, \mathcal{D}\phi_{k'} \rangle = \kappa^2 \langle \psi_k, \phi_{k'} \rangle + \left\langle \frac{\partial}{\partial t} \psi_k, \frac{\partial}{\partial t} \phi_{k'} \right\rangle - \langle \psi_k, \partial_n \phi_{k'} \rangle_{\partial \bar{T}}. \quad (27)$$

Table 11. Algorithm 1: group formation for the grouped subsampler

1	<i>procedure</i> Group-formation ($\mathbf{x}_1, \dots, \mathbf{x}_m$)	
2	$\mathcal{I} \leftarrow \{1, \dots, m\}$	m is the total number of subjects to group
3	$k \leftarrow 1$	
4	$\mathcal{G}_0 \leftarrow \emptyset$	
5	<i>while</i> $ \mathcal{I} > 0$ <i>do</i>	
6	$\tilde{\mathcal{G}} \leftarrow \text{Create-group}(\mathcal{I}, \mathbf{x}_1, \dots, \mathbf{x}_m)$	
7	<i>if</i> $\text{rank}(\sum_{i \in \tilde{\mathcal{G}}} \mathbf{x}_i \mathbf{x}_i^T) = \text{columns}(\mathbf{x}_1)$ <i>then</i>	$\text{columns}(\mathbf{x}_1)$ is the number of covariates
8	$\mathcal{G}_k \leftarrow \tilde{\mathcal{G}}$	
9	$\mathcal{I} \leftarrow \mathcal{I} \setminus \mathcal{G}_k$	
10	$k \leftarrow k + 1$	
11	<i>else</i>	
12	$\mathcal{G}_0 \leftarrow \mathcal{I}$	
13	$\mathcal{I} \leftarrow \emptyset$	
14	<i>end if</i>	
15	<i>end while</i>	
16	<i>return</i> $\mathcal{G}_0, \dots, \mathcal{G}_k$	
17	<i>end procedure</i>	
18	<i>procedure</i> Create-group($\mathcal{I}, \mathbf{x}_1, \dots, \mathbf{x}_m$)	
19	$\mathcal{G} \leftarrow \mathcal{I}_1$	
20	$\mathcal{I} \leftarrow \mathcal{I} \setminus \mathcal{I}_1$	
21	<i>while</i> $\text{rank}(\sum_{i \in \mathcal{G}} \mathbf{x}_i \mathbf{x}_i^T) < \text{columns}(\mathbf{x}_1)$ <i>and</i> $ \mathcal{I} > 0$ <i>do</i>	
22	<i>if</i> $\text{rank}(\mathbf{x}_{\mathcal{I}_1} \mathbf{x}_{\mathcal{I}_1}^T + \sum_{i \in \mathcal{G}} \mathbf{x}_i \mathbf{x}_i^T) > \text{rank}(\sum_{i \in \mathcal{G}} \mathbf{x}_i \mathbf{x}_i^T)$ <i>then</i>	
23	$\mathcal{G} \leftarrow \mathcal{G} \cup \mathcal{I}_1$	
24	<i>end if</i>	
25	$\mathcal{I} \leftarrow \mathcal{I} \setminus \mathcal{I}_1$	
26	<i>end while</i>	
27	<i>return</i> \mathcal{G}	
28	<i>end procedure</i>	

Here the final term vanishes if Dirichlet or Neumann boundary conditions are used.

For the NIG version of the model, we approximate the distribution of L_k by

$$L_k = h_k \delta^W + \mu^W V_k^W + \sqrt{V_k^W} Z_k,$$

where $Z_k \sim N(0, 1)$, $h_k = \langle \psi_k, 1 \rangle$ and $V_k \sim \text{IG}(\nu, h_k^2 \nu)$ (Bolin, 2014). It follows that the distribution for the stochastic weight vector \mathbf{W} conditional on V can be written as equation (15).

When $\mathcal{D} = \kappa + \partial/\partial t$ is a first-order operator, we cannot use the Galerkin method. We then instead use a Petrov–Galerkin method, where the test functions above are replaced by piecewise constant functions:

$$\psi_i(t) = \begin{cases} 1, & s_i < t < s_{i+1}, \\ 0, & \text{otherwise.} \end{cases}$$

With this change, the distribution of \mathbf{L} is the same as above (which is not an approximation in this case), but the elements of \mathbf{K} are

$$K_{kk'} = \langle \psi_k, \mathcal{D} \phi_{k'} \rangle = \kappa \langle \psi_k, \phi_{k'} \rangle + \left\langle \psi_k, \frac{\partial}{\partial t} \phi_{k'} \right\rangle. \quad (28)$$

If the operator is an integer power of a first- or second-order operator, the model can be rewritten as a system of equations. For example $\mathcal{D}^2 W(t) = dL(t)$ can be formulated as the system

$$\begin{aligned} \mathcal{D}W(t) &= u(t), \\ \mathcal{D}u(t) &= dL(t). \end{aligned}$$

Both of these equations can then be discretized by using the method above. Combining the two discretizations yields the following equation for the coefficients, $\mathbf{KCKW} = \mathbf{L}$, where \mathbf{C} is the mass matrix with elements $C_{kk'} = \langle \psi_k, \phi_{k'} \rangle$. If the operator is a fractional power of a first- or second-order operator, the iterative formulation cannot be used. However, the fractional power could probably still be handled by using the methods in Bolin *et al.* (2020) and Bolin and Kirchner (2020).

References

- Andrieu, C., Moulines, É. and Priouret, P. (2007) Stability of stochastic approximation under verifiable conditions. *SIAM J. Control Optimizn.*, **44**, 283–312.
- Aralleno-Valle, R. B., Bolfarine, H. and Lachos, V. H. (2007) Bayesian inference for skew-Normal linear mixed models. *J. Appl. Statist.*, **34**, 663–682.
- Asar, Ö., Ritchie, J. P., Kalra, P. A. and Diggle, P. J. (2016) Short-term and long-term effects of acute kidney injury in chronic kidney disease patients: a longitudinal analysis. *Biomet. J.*, **58**, 1552–1566.
- Bai, X., Chen, K. and Yao, W. (2016) Mixture of linear mixed models using multivariate t distribution. *J. Statist. Computn Simuln.*, **86**, 771–787.
- Barndorff-Nielsen, O. E. (1977) Exponentially decreasing distributions for the logarithm of the particle size. *Proc. R. Soc. A*, **353**, 401–419.
- Barndorff-Nielsen, O. (1997a) Processes of normal inverse Gaussian type. *Finan. Stochast.*, **2**, 41–68.
- Barndorff-Nielsen, O. (1997b) Normal inverse Gaussian distributions and stochastic volatility modelling. *Scand. J. Statist.*, **24**, 1–13.
- Bibby, B. and Sørensen, M. (2003) Hyperbolic processes in Finance. In *Handbook of Heavy Tailed Distributions in Finance* (ed. S. T. Rachov), pp. 211–248. Amsterdam: North-Holland.
- Bolin, D. (2014) Spatial Matérn fields driven by non-Gaussian noise. *Scand. J. Statist.*, **41**, 557–579.
- Bolin, D. and Kirchner, K. (2020) The rational SPDE approach for Gaussian random fields with general smoothness. *J. Computnl Graph. Statist.*, **29**, 274–285.
- Bolin, D., Kirchner, K. and Kovacs, M. (2020) Numerical solution of fractional elliptic stochastic PDEs with spatial white noise. *IMA J. Numer. Anal.*, **40**, 1051–1073.
- Bolin, D. and Wallin, J. (2020) Multivariate type G Matérn stochastic partial differential equation random fields. *J. R. Statist. Soc. B*, **82**, 215–239.
- Cabral, C. R., Lachos, V. H. and Madruga, M. R. (2012) Bayesian analysis of skew-Normal independent linear mixed models with heterogeneity in the random-effects population. *J. Statist. Planng Inf.*, **142**, 181–200.
- Choudhary, P. K., Sengupta, D. and Cassey, P. (2014) A general skew- t mixed model that allows different degrees of freedom for random effects and error distribution. *J. Statist. Planng Inf.*, **147**, 235–247.
- Davidian, M. and Gallant, A. R. (1993) The nonlinear mixed effects models with a smooth random effects density. *Biometrika*, **80**, 475–488.

- Davies, J. C. and Alton, E. W. (2009) Monitoring respiratory disease severity in cystic fibrosis. *Resp. Med.*, **54**, 606–617.
- De la Cruz, R. (2014) Bayesian analysis for nonlinear mixed-effects models under heavy-tailed distributions. *Pharmaceut. Statist.*, **13**, 81–93.
- Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. B*, **39**, 1–38.
- Diggle, P. J. (1988) An approach to the analysis of repeated measurements. *Biometrics*, **44**, 959–971.
- Diggle, P. J., Heagerty, P. J., Liang, K.-Y. and Zeger, S. L. (2002) *Analysis of Longitudinal Data*, 2nd edn. Oxford: Oxford University Press.
- Diggle, P. J., Sousa, I. and Asar, Ö. (2015) Real-time monitoring of progression towards renal failure in primary care patients. *Biostatistics*, **16**, 522–536.
- Eberlein, E. (2001) Application of generalized hyperbolic Lévy motions to finance. In *Lévy Processes: Theory and Applications* (eds O. E. Barndorff-Nielsen, S. I. Resnick and T. Mikosch), pp. 319–337. Boston: Birkhäuser.
- Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2011) *Applied Longitudinal Analysis*, 2nd edn. Hoboken: Wiley.
- Ghidey, W., Lesaffre, E. and Eilers, P. (2004) Smooth random effects distribution in a linear mixed model. *Biometrics*, **60**, 945–953.
- Henderson, R., Diggle, P. and Dobson, A. (2000) Joint modelling of longitudinal measurements and event time data. *Biostatistics*, **1**, 465–480.
- Ho, H. L. and Lin, T.-I. (2010) Robust linear mixed models using the skew t distribution with application to schizophrenia data. *Statist. Med.*, **52**, 449–469.
- Jara, A., Quintana, F. and Martín, E. S. (2008) Linear mixed models with skew-elliptical distributions: a Bayesian approach. *Computnl Statist. Data Anal.*, **52**, 5033–5045.
- Jennrich, R. I. and Schluchter, M. D. (1986) Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, **42**, 805–820.
- Jørgensen, B. (1982) *Statistical Properties of the Generalized Inverse Gaussian Distribution*, pp. 401–419. Berlin: Springer.
- Kay, S. R., Fiszbein, A. and Opler, L. A. (1987) The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. *Schiz. Bull.*, **13**, 261–276.
- Kazemi, I., Mahdizadeh, Z., Mansourian, M. and Park, J. J. (2013) Bayesian analysis of multivariate mixed models for a prospective cohort study using skew-elliptical distributions. *Biometr. J.*, **55**, 495–508.
- Kleinman, K. P. and Ibrahim, J. G. (1998) A semiparametric Bayesian approach to the random effects model. *Biometrics*, **54**, 921–938.
- Koller, M. (2016) robustlmm: an R package for robust estimation of linear mixed-effects models. *J. Statist. Softw.*, **75**, 1–24.
- Koller, M. and Stahel, W. A. (2016) Nonsingular subsampling for regression S estimators with categorical predictors. *Computnl Statist.*, **32**, 1–16.
- Kollo, T. and von Rosen, D. (2006) Advanced multivariate statistics with matrices (ed. M. Hazewinkel). In *Mathematics and Its Applications*. New York: Springer.
- Kushner, H. and Yin, G. (2003) *Stochastic Approximation and Recursive Algorithms and Applications*. Berlin: Springer.
- Lachos, V. H., Bandyopadhyay, D. and Dey, D. K. (2011) Linear and nonlinear mixed-effects models for censored HIV viral loads using Normal/independent distributions. *Biometrics*, **67**, 1594–1604.
- Lachos, V. H., Cabral, C. R. B. and Abanto-Valle, C. A. (2012) A non-iterative sampling Bayesian method for linear mixed models with Normal independent distributions. *J. Appl. Statist.*, **39**, 531–549.
- Lachos, V. H., Castro, L. M. and Dey, D. K. (2013) Bayesian inference in nonlinear mixed-effects models using Normal independent distributions. *Computnl Statist. Data Anal.*, **64**, 237–252.
- Lachos, V. H., Dey, D. K. and Cancho, V. G. (2009) Robust linear mixed models with skew-Normal independent distributions from a Bayesian perspective. *J. Statist. Plannng Inf.*, **139**, 4098–4110.
- Lachos, V. H., Ghosh, P. and Arellano-Valle, R. B. (2010) Likelihood based inference for skew-Normal independent linear mixed models. *Statist. Sin.*, **20**, 302–322.
- Laird, N. M. and Ware, J. H. (1982) Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.
- Lange, K. (1995) A gradient algorithm locally equivalent to the EM algorithm. *J. R. Statist. Soc. B*, **57**, 425–437.
- Lange, K. L., Little, R. J. A. and Taylor, J. M. G. (1989) Robust statistical modeling using the t distribution. *J. Am. Statist. Ass.*, **84**, 881–896.
- Lange, K. and Sinsheimer, J. S. (1993) Normal/independent distributions and their applications in robust regression. *J. Computnl Graph. Statist.*, **2**, 175–198.
- Levey, A. S., Bosch, J. P., Lewis, J. B., Greene, T., Rogers, N. and Roth, D. (1999) A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. *Ann. Intern. Med.*, **130**, 461–470.
- Liang, K.-Y. and Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.
- Lin, T. I. and Lee, J. C. (2007) Bayesian analysis of hierarchical linear mixed modeling using the multivariate t distribution. *J. Statist. Plannng Inf.*, **137**, 484–495.

- Lin, T.-I. and Wang, W. L. (2011) Bayesian inference in joint modelling of location and scale parameters of the t distribution for longitudinal data. *J. Statist. Plannng Inf.*, **141**, 1543–1553.
- Lin, T.-I. and Wang, W. L. (2013) Multivariate skew-Normal linear mixed models for multi-outcome longitudinal data. *Statist. Modelling*, **13**, 199–221.
- Lindgren, F. and Rue, H. (2008) On the second-order random walk model for irregular locations. *Scand. J. Statist.*, **35**, 691–700.
- Lindgren, F., Rue, H. and Lindström, J. (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach (with discussion). *J. R. Statist. Soc. B*, **73**, 423–498.
- Liu, C. and Rubin, D. B. (1995) ML estimation of the t distribution using EM and its extensions, ECM and ECME. *Statist. Sin.*, **5**, 19–39.
- Louis, T. A. (1982) Finding the observed information matrix when using the EM algorithm. *J. R. Statist. Soc. B*, **44**, 226–233.
- Lu, Z. and Zhang, Z. (2014) Robust growth mixture models with non-ignorable missingness: models, estimation, selection, and application. *Computnl Statist. Data Anal.*, **71**, 220–240.
- Magnus, J. R. and Neudecker, H. (1979) The commutation matrix: some properties and applications. *Ann. Statist.*, **7**, 381–394.
- Magnus, J. R. and Neudecker, H. (2007) *Matrix Differential Calculus with Applications in Statistics and Econometrics*, 3rd edn. Chichester: Wiley.
- Matérn, B. (1960) *Spatial Variation*. Stockholm: Statens Skogsforsningsinstitut.
- Matos, L. A., Prates, M. O., Chen, M.-H. and Lachos, V. H. (2013) Likelihood-based inference for mixed-effects models with censored response using the multivariate- t distribution. *J. Computnl Graph. Statist.*, **10**, 249–276.
- Meza, C., Osorio, F. and De la Cruz, R. (2012) Estimation in nonlinear mixed-effects models using heavy-tailed distributions. *Statist. Comput.*, **22**, 121–139.
- Molenberghs, G. and Verbeke, G. (2005) *Models for Discrete Longitudinal Data*. New York: Springer.
- Osorio, F. (2016) heavy: robust estimation using heavy-tailed distributions. *R Package Version 0.3*. (Available from <http://cran.r-project.org/package=heavy>.)
- Pinheiro, J. C., Liu, C. and Wu, Y. N. (2001) Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *J. Computnl Graph. Statist.*, **10**, 249–276.
- Podgórski, K. and Wallin, J. (2016) Convolution-invariant subclasses of generalized hyperbolic distributions. *Commun Statist. Theory Meth.*, **45**, 98–103.
- Rizopoulos, D. (2012) *Joint Models for Longitudinal and Time-to-event Data: with Applications in R*. Boca Raton: Chapman and Hall–CRC.
- Rosa, G. J. M., Gianola, D. and Padovani, C. R. (2004) Bayesian longitudinal data analysis with mixed models and thick-tailed distributions using MCMC. *J. Appl. Statist.*, **31**, 855–873.
- Rosa, G. J. M., Padovani, C. R. and Gianola, D. (2003) Robust linear mixed models with Normal/independent distributions and Bayesian MCMC implementation. *Biometr. J.*, **45**, 573–590.
- Song, P. X.-K., Zhang, P. and Qu, A. (2007) Maximum likelihood inference in robust linear mixed-effects linear mixed effects models using multivariate t distributions. *Statist. Sin.*, **17**, 929–943.
- Stirrup, O. T., Babiker, A. G., Carpenter, J. R. and Copas, A. J. (2015) Fractional Brownian motion and multivariate- t models for longitudinal biomedical data, with application to CD4 counts in HIV-patients. *Statist. Med.*, **35**, 1514–1532.
- Subtil, F. and Rabilloud, M. (2010) Robust non-linear mixed modelling of longitudinal PSA levels after prostate cancer treatment. *Statist. Med.*, **29**, 573–587.
- Sun, J., Frees, E. W. and Rosenberg, M. A. (2008) Heavy-tailed longitudinal modeling using copulas. *Insur. Math. Econ.*, **42**, 817–830.
- Tankov, P. (2003) *Financial Modelling with Jump Processes*. Boca Raton: Chapman and Hall–CRC.
- Tao, H., Palta, M., Yandell, B. S. and Newton, M. A. (2004) An estimation method for the semiparametric mixed effects model. *Biometrics*, **55**, 102–110.
- Taylor, J. M. G., Cumberland, W. G. and Sy, J. P. (1994) A stochastic process model for analysis of longitudinal AIDS data. *J. Am. Statist. Ass.*, **89**, 727–736.
- Taylor-Robinson, D., Whitehead, M., Diderichsen, F., Olesen, H. V., Pressler, T., Smyth, R. L. and Diggle, P. J. (2012) Understanding the natural progression in %FEV decline in patients with cystic fibrosis: a longitudinal study. *Thorax*, **67**, 860–866.
- Tian, G.-L., Ng, K. W. and Tan, M. (2008) EM-type algorithms for computing restricted MLEs in multivariate Normal distributions and multivariate t -distributions. *Computnl Statist. Data Anal.*, **52**, 4768–4778.
- Verbeke, G. and Lesaffre, E. (1996) A linear mixed-effects model with heterogeneity in the random-effects population. *J. Am. Statist. Ass.*, **91**, 217–221.
- Verbeke, G. and Molenberghs, G. (2001) *Linear Mixed Models for Longitudinal Data*. New York: Springer.
- Vilca, F., Balakrishnan, N. and Zeller, C. B. (2014) Multivariate skew-Normal generalized hyperbolic distribution and its properties. *J. Multiv. Anal.*, **128**, 74–85.
- Vock, D. M., Davidian, M. and Tsiatis, A. A. (2012) Mixed model analysis of censored longitudinal data with flexible random-effects density. *Biostatistics*, **13**, 61–73.

- Wahba, G. (1990) *Spline Models for Observational Data*. Philadelphia: Society for Industrial Mathematics.
- Wallin, J. and Bolin, D. (2015) Geostatistical modelling using non-Gaussian Matérn fields. *Scand. J. Statist.*, **42**, 872–890.
- Wang, W.-L. and Fan, T.-H. (2011) Estimation in multivariate t linear mixed models for multiple longitudinal data. *Statist. Sin.*, **21**, 1857–1880.
- Wang, W.-L. and Fan, T.-H. (2012) Bayesian analysis of multivariate t linear mixed models using a combination of IBF and Gibbs sampler. *J. Multiv. Anal.*, **105**, 300–310.
- Wang, W.-L., Lin, T.-I. and Lachos, V. H. (2015) Extending multivariate- t linear mixed models for multiple longitudinal data with censored responses and heavy tails. *Statist. Meth. Med. Res.*, **27**, 48–69.
- Yavuz, F. G. and Arslan, O. (2016) Linear mixed model with Laplace distribution (LLMM). *Statist. Pap.*, **59**, 271–289.
- Zeller, C. B., Labra, F. V., Lachos, V. H. and Balakrishnan, N. (2010) Influence analyses of skew-Normal/independent linear mixed models. *Computnl Statist. Data Anal.*, **54**, 1266–1280.
- Zhang, D. and Davidian, M. (2001) Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, **57**, 795–802.
- Zhang, J., Yu, B., Zhang, L., Roskos, L., Richman, L. and Yang, H. (2015) Non-Normal random effects models for immunogenicity assay cut point determination. *J. Biopharm. Statist.*, **25**, 295–306.
- Zhang, P., Qui, Z., Fu, Y. and Song, P. X.-K. (2009) Robust transformation mixed-effects models for longitudinal continuous proportional data. *Can. J. Statist.*, **37**, 266–281.
- Zhu, B. and Dunson, D. B. (2017) Bayesian functional data modeling for heterogeneous volatility. *Baysn Anal.*, **12**, 335–350.
- Zhu, B., Song, P. X.-K. and Taylor, J. M. G. (2011a) Stochastic functional data analysis: a diffusion model-based approach. *Biometrics*, **67**, 1295–1304.
- Zhu, B., Taylor, J. M. G. and Song, P. X.-K. (2011b) Semiparametric stochastic modeling of the rate function in longitudinal studies. *J. Am. Statist. Ass.*, **106**, 1485–1495.

Discussion on the paper by Asar, Bolin, Diggle and Wallin

Arnošt Komárek (Charles University, Prague)

The approach of Asar, Bolin, Diggle and Wallin is motivated by a classical linear mixed effects model (Laird and Ware, 1982) for repeated or longitudinal outcomes $Y_{i,j}$, $i = 1, \dots, m$, $j = 1, \dots, n_i$, that can be specified as

$$Y_{i,j} = \mathbf{x}_{i,j}^T \boldsymbol{\beta} + \mathbf{d}_{i,j}^T \mathbf{U}_i + \sigma Z_{i,j}, \quad (29)$$

where $\mathbf{x}_{i,j}$ and $\mathbf{d}_{i,j}$ are vectors of covariates, usually involving time $t_{i,j}$ of measurements. Further, $\boldsymbol{\beta}$ and σ are unknown parameters, \mathbf{U}_i is the random-effects vector and $Z_{i,j}$ is the error term of the model. It is traditionally assumed that both random effects and the error term are independent and identically distributed and Gaussian. As Asar and his colleagues state, a (perhaps too) frequently used variant of model (29) is a simple ‘random-intercept and random-slope’ model where $\mathbf{d}_{i,j}^T \mathbf{U}_i \equiv U_{i,1} + U_{i,2} t_{i,j}$. This is clearly inappropriate in many practical settings. To overcome this model deficiency, they follow earlier proposals and add a stochastic process $W_i(t_{i,j})$, specified through a parametric covariance function, to the model formula (29). They mention that an alternative way to capture non-linear behaviour of repeated measurements is to use, for example, regression splines instead of the stochastic process, i.e. the process $W_i(t_{i,j})$ would be replaced by a term like

$$\sum_{k=1}^K U_{i,k} B_k(t_{i,j}) \quad (30)$$

with $\mathbf{U}_i = (U_{i,1}, \dots, U_{i,K})^T$ a vector of random effects and $B_1(t), \dots, B_K(t)$ a suitable spline basis. In my opinion, the two approaches are complementary rather than competing. The stochastic process approach is certainly more parsimonious with respect to the number of unknown parameters and, as such, it would probably be the first choice if prediction is the primary aim of the analysis. However, if the main interest lies in modelling the shape of both the subject-specific and the mean trajectory then regression splines seem more natural to me and enable an easier interpretation of the fitted model by practitioners.

Not only the assumed shape of a (subject-specific) trajectory (model formula (29)) but also distributional assumptions make the model either only wrong or wrong but useful. As mentioned above and also by Asar and his colleagues, all random components of the model, distribution of the random effects \mathbf{U}_i , the error terms $Z_{i,j}$ and possible additional model components, e.g. the stochastic process $W_i(t)$, are often assumed

to be Gaussian. It is clear that it is a convolution of those random elements which determines the marginal distribution of the outcome vector $\mathbf{Y}_i = (Y_{i,1}, \dots, Y_{i,n})^T$, enters the likelihood and is identifiable from data. Hence it is not possible to be fully flexible ('non-parametric') in the specification of all the random components. Asar and his colleagues propose a generalized inverse Gaussian (GIG) distributional class which is still fully parametric, yet quite flexible. An algorithm to calculate the maximum likelihood estimate is not only proposed but also implemented along with additional supplementary routines in a form of the contributed R package *ngme*.

Here, I should like to make two proposals for possible future development of the model of Asar and his colleagues. In all applications in Section 6, only the random intercept, i.e. univariate random effect U_i , is considered next to the stochastic process $W_i(t)$ to capture non-linearity in subject-specific trajectories. Suppose that someone prefers a more structured specification of the subject-specific trajectories made by, for example, the regression splines (30) instead of the stochastic process. This would clearly require a multivariate random-effect vector \mathbf{U}_i and it would be interesting to see how the estimation methods perform in this case. Second, the GIG distributional class is in fact specified hierarchically (equation (10)). Hence the GIG class seems to be perfectly suited for use within Bayesian models and inference based on Markov chain Monte Carlo and related methods. As not everybody would want to use Bayesian approaches, e.g. because priors must be specified. I appreciate the maximum likelihood estimation based procedures of Asar and his colleagues. However, MCMC-based Bayesian inference seems easier and perhaps enables use of the GIG class within models of even a higher complexity.

In summary, I congratulate the authors for an excellent paper with great potential to trigger further research and development of additional 'wrong but useful' models for increasingly complex data sets that are continuously appearing in very different research areas. Last, but not the least, the software implementation of the methodology should be applauded as it highly increases the chance that the method will also be used to solve real problems.

Daniel Farewell (*Cardiff University*)

I am grateful for the invitation to study this paper and to pose questions to its authors. The paper exhibits their collective and characteristic skill in addressing substantive clinical questions like 'is this patient losing kidney function at a relative rate of at least 5%?', which they answer by using nowcasting based on estimated random effects. My personal highlight is their demonstration that normal-inverse Gaussian modelling markedly influences this answer (Fig. 9). I remember reading Verbeke and Lesaffre (1997) and extrapolating (quite unjustifiably, as I now see) that I need never fret over whether random effects might deviate from normality. That only fixed effect estimation is reasonably unaffected by such departures is an important subtlety to discussions of robustness that is presented particularly clearly here.

I also want to commend the authors' graphical presentation of both data and model diagnostics. In comparison with the customary spaghetti plots that are ubiquitous in exploratory analysis of longitudinal data. I found the 'spaghetti-and-meatballs' plots that are favoured by the authors much more helpful: they clearly illustrate both the measurement range and some carefully selected individual dynamics of subject-specific trajectories, with neither obscuring the other.

A third point of interest was the authors' use of the duality between Gaussian processes expressed in terms of covariance functions and as solutions to stochastic differential equations. The latter dynamic perspective is appealing for the study of longitudinal data, but it is still reasonably rare (see Pan and Mackenzie (2003) for an exception). I was pleased that some of the authors' models were *not* time reversible, and I invite them to consider walking even further down this road, making time a fundamental and irreversible feature.

A dynamic approach simplifies the treatment of informative observation or dropout: as an example, the authors mention patients who have many measurements during periods of illness. By contrast, conditions like missingness at random are purely probabilistic and agnostic to temporal ordering. This makes it difficult to assess whether an observation mechanism is ignorable. The standard approach to longitudinal data analysis is in the same spirit: begin with a 'complete, underlying' longitudinal trajectory $(Y_i : t \in \mathbb{N} \text{ or } \mathbb{R}_+)$ for each subject, which is then incompletely observed at times t_1, t_2 and so on. Yet a quantity like %FEVI cannot be defined in continuous time: by definition, it is an average over a time period of 1 s, and more importantly exhalation cannot be sustained indefinitely! Any continuous time quantification of %FEVI must therefore involve so-called cross-world counterfactuals (Richardson and Robins, 2013), and modelling the covariance of cross-world counterfactuals is best avoided wherever possible. Perhaps more importantly, the standard missing data approach struggles under the weight of the static, global perspective that it induces, whereas a dynamic approach can more easily be melded to the substantive problem at hand (Farewell *et al.*, 2017).

One possible alternative is to define the continuous time process of interest not as an ‘underlying’ measurement of %FEVI (say), but instead as local characteristics of the data themselves. The data arising from a given subject may be represented as a marked point process $\{(T_j, Y_j) : j = 1, 2, 3, \dots\}$, where T_j records the time that measurement j is taken, and Y_j the corresponding value of that measurement (Jacobsen, 2006). As in the Doob–Meyer decomposition, this stochastic process may be partitioned into ‘signal’ and ‘noise’, the signal being a previsible process and the noise a martingale. I would like to know whether the authors think that they could rephrase practical questions like ‘are they losing function at a rate of 5%’ with reference to the intensity of a marked point process rather than an ‘underlying’ variable?

Contrast this decomposition with the components of variance that are described by the authors’ model. Their first separation into ‘signal’ and ‘noise’ isolates U and W from the ‘measurement error’ Z . Again, this invites speculation about and perhaps even presupposes a hypothetical error-free ‘measurement’ that exists at every time point. But is this unambiguously defined? It is a little like asking whether there is (in principle) an error-free photograph of a beautiful landscape: the subject matter is massively multivariate, and defining a suitably pure projection of the phenomenon onto a two-dimensional representation is challenging. The underlying principle of the authors’ modelling seems to be that this pure, low dimensional representation can be defined by removing high frequency components like Z . Do the authors agree and, if so, do they see merit in a more formal use of frequency domain decompositions?

I am also curious about the second decomposition into U and W . Would the authors contend that these can be interpreted separately, and are there prediction tasks for which they feel that such a separation might be important?

Gertrude Stein noted that ‘[s]ilent gratitude isn’t very much use to anyone’. In seconding the vote of thanks, I therefore declare my loud admiration for this paper, and I invite the audience to join me in expressing gratitude to the authors for their beautiful and practical research.

The vote of thanks was passed by acclamation.

Alex Stringer and Patrick Brown (*University of Toronto and Centre for Global Health Research, Toronto*) and **Jamie Stafford** (*University of Toronto*)

We congratulate Asar, Bolin, Diggle and Wallin on an important development in an important topic and believe that their new model and software will become an important part of the statistician’s toolbox. We especially appreciate the use of stochastic optimization for maximizing the complicated marginal likelihood that arises from their procedure. In principle, *any* non-convex optimization algorithm could be used. In practice, stochastic optimization algorithms are popular in the machine learning literature for optimizing complicated non-convex objective functions and we believe that they should be more widely considered in statistical applications when appropriate.

In our experience, stochastic optimization becomes more appealing as the size of the data set increases and the objective function becomes more complex. The authors’ use of several optimizations running in parallel suggests that computation time and memory use are of lesser concern than stability in their application. Would they comment on the specific factors that motivate their use of stochastic optimization, and the effect that it had on computation time and memory use? Under what circumstances would multiple subsampling-based optimizations in parallel be preferable to running a single optimization using the full data set? Out of the box procedures are available in R (e.g. the `ipoptR` software) that are highly stable and efficient, and it would be interesting to assess their performance on this interesting and complex problem.

A related point is that the authors choose to develop several custom procedures, specifically for assessing convergence, preconditioning and ensuring that intermediate quantities remain computable during the optimization. Have they considered using an out of the box method for gradient-based stochastic optimization? There is a rich literature on this topic in the field of machine learning. For example, the Adam algorithm (Kingma and Ba, 2015) effectively uses a bias-corrected gradient-based preconditioner that we have found to be fast, memory efficient and stable in large-scale problems. This would eliminate the need for the grouped subsampling procedures of Section 4.4 and potentially improve the computation time of the authors’ procedure.

Neil K. Chada and Ajay Jasra (*King Abdullah University of Science and Technology, Thuwal*)

We thank Asar, Bolin, Diggle and Wallin for this thought-provoking paper, which is concerned with understanding and introducing a non-Gaussian phenomenon in mixed models of the form

$$Y_{ij} = +x_{ij}^T \beta + d_{ij}^T U_i + W_i(t_{ij}) + \alpha Z_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, m, \quad (31)$$

where $U_i \sim N(0, \Sigma)$, $W_i(t)$ and $Z_{ij} \sim N(0, 1)$ are stochastic components and β is a parameter of interest.

This motivation is to account for the non-linear trajectories of the stochastic effects within these models. Traditionally much of the work in this field has assumed that these components are normally distributed. The authors consider tackling this by presenting the multivariate generalized hyperbolic distribution, which in itself includes many widely used distributions such as the generalized asymmetric Laplace, Student t - and Cauchy distributions. To introduce the non-Gaussian stochastic process $W_i(t)$ of model (31) they solve a stochastic differential equation

$$\mathcal{D}W_i(t) = dL_i(t).$$

such that \mathcal{D} is a particular choice of an operator and $dL_i(t)$ is a non-Gaussian Lévy process. This links well with previously used methodologies in that solving a particular stochastic partial differential equation will lead to a Gaussian random field with a Matérn covariance function, with the change that $dL_i(t)$ represents Gaussian white noise (Lindgren *et al.*, 2011).

A recent trend in the field of statistical inverse problems (Stuart, 2010), which is concerned with parameter estimation, has been incorporating non-Gaussian prior information. Some of this work (Chada *et al.*, 2019) has focused on Cauchy processes which arise from α -stable distributions. If we assume that $W_i(t)$ is a random field defined by an α -stable distribution, then it can be expressed as

$$W_i(t) \sim S_\alpha(\mu, \beta, \sigma),$$

where $\alpha \in (0, 2]$ denotes the stability parameter and $\mu \in (-\infty, \infty)$, $\beta \in [-1, 1]$ and $\sigma \in [0, \infty)$ denote the location, skewness and scale parameters respectively. Chada *et al.* (2019) provided theoretical justifications to promote the use of these fields, as well as particular forms of discretization. Other related work has been on choosing the hyperparameters (Roininen *et al.*, 2019) of the covariance function through non-Gaussian distributions.

We believe that the work in Chada *et al.* (2019) can be potentially used in the context of the authors. This would enable us to model the random fields through generalized hyperbolic distributions which may provide closed form expressions. This could also be done through the stochastic partial differential equation formulation where further probabilistic understanding could be addressed, such as Wiener's Tauberian theory. Finally connections could be made through the various approaches to construct Cauchy processes.

Christine P. Chai (*Microsoft Corporation, Redmond*)

I thank Asar, Bolin, Diggle and Wallin for improving the linear mixed effects model by adding a subject-specific continuous time stochastic process $W_i(t)$, which is important because each subject reacts differently over time. The model validation and model selection section is short but excellently written! The section successfully explains why using a normal $Q-Q$ -plot would be insufficient to identify the source of non-Gaussianity and points out that the random-effect component U_i would interfere with the $Q-Q$ -plot results of Y_{ij} . It is a better approach to compare the prediction of each model component with the simulated data, which avoids the distortion from correlation over time.

Since the repeated measurement data are assumed to be continuous, my major question would be: how are missing data handled in the model? Was any interpolation or extrapolation used? For example, a patient may have measurements from time t_1 to t_k except for t_3 and t_5 . Another patient may be missing t_2 and t_{k-1} . It would be really impressive if neither of the two data sets in the case-studies contains any missing data.

Lastly, I mention a minor issue: the normal-inverse Gaussian model has better predictive power than the Gaussian model, but this does not imply that it is 'correct'. As Box (1976) said, 'All models are wrong, but some are useful'.

(The opinions and views expressed here are those of the author and do not necessarily state or reflect those of Microsoft.)

Kuldeep Kumar (*Bond University, Gold Coast*)

I congratulate Asar, Bolin, Diggle and Wallin for proposing a novel method to analyse continuous repeated measure outcomes collected longitudinally and for enabling a combination of stochastic components to be non-Gaussian. This kind of data is quite often encountered not only in medical sciences but also in social sciences when repeated measurements of an outcome variable are made on subjects with follow-up over time, which may vary from subject to subject. The authors consider baseline covariates as well as longitudinal covariates and propose a methodology to estimate parameters for the data, which have departed skewed from the Gaussian distributional assumption. They basically propose two procedures for the estimation of parameters: stochastic gradient, and multiple-chain estimation. My trifling disappointment is that they

did not deal with measurement error in longitudinal covariates, which are quite often encountered in social science data. In linear mixed effects models, the regularization methods using penalized profile likelihood have been developed to select and estimate fixed and random effects. These methods avoid the stochastic error of variable selection in stepwise procedures and can successfully handle the high dimensional setting in both fixed and random effects, as cited by Fan and Li (2012). Such penalized profile likelihood methods can be developed for the general specification of linear mixed effects models for repeated measurement data.

The following contributions were received in writing after the meeting.

Anna L. Choi and Tze Leung Lai (*Stanford University*)

This paper presents an innovative method to assess each of the three stochastic components independently for non-Gaussian distributions in a linear mixed model, when one or more of the components do not satisfy the Gaussian assumption. It uses an elegant and computationally efficient subsampling-based stochastic gradient algorithm to develop a novel implementation to meet the computational challenges of large data sets, together with an R package *ngame*. Two case-studies involving reanalysis of cystic fibrosis data and renal data are used to compare the Gaussian and non-Gaussian models, with the latter data set giving different results and showing the advantages of the proposed assessment or diagnosis method. The last paragraph of the discussion section mentions possible extension of the methodology and the software package to spatial and multivariate models, for both the longitudinal and the classical geostatistical settings. We have followed up on this lead in an on-going project to forecast Covid-19 transmissions and have found it very useful.

Emrah Gecili and Rhonda Szczesniak (*University of Cincinnati College of Medicine*)

We applaud Asar, Bolin, Diggle and Wallin for their outstanding presentation during the discussion session and highlighting important clinical research ramifications of their paper. Describing the time course of cystic fibrosis (CF) lung function was a compelling application, justifying the importance of introducing and relaxing assumptions of the process $W_i(t_{ij})$ and measurement error Z_{ij} . By framing the prediction problem as a target function, one can use their model to form predictive probability distributions of rapid disease progression and translate probabilities for downstream informatics (Szczesniak *et al.*, 2019). In Fig. 12 showing an R Shiny web application, the target function varies from -10% to 0.5% predicted/year, reflecting lung function trajectory and rate of change in A–B (see the delta threshold, circled; the predictive probability over time in C changes according to the selected target value). Patient covariate inputs are shown at the left-hand side and in D (www.predictfev1.com). Further bolstering the applied case for this method is its potential utility in biomarker-informed prediction modelling. As noted by Dr Kumar during the meeting, measurement error in longitudinal covariates was not addressed; however, this could be ameliorated (or at least these errors could be shrunk) by using regularization. Implementing different types of regularizations for this model could be facilitated in the Bayesian framework by using various shrinkage priors, thereby enabling simultaneous variable selection to be performed with the model proposed. This could provide a useful tool for researchers from different areas of applied science. These penalized regression models can accommodate the stochastic error of variable-selection stepwise procedures (mentioned by Dr Kumar) and easily handle high dimensional variable-selection problems, which are commonly observed in modern science. Of course, advantages of using Bayesian regularized methods would come with the cost of increased computational time, especially for the cases where the sample size is large. Specific to the CF case, we modelled longitudinal lung function data by using an earlier (Gaussian) iteration of work by a subset of the authors (Diggle *et al.*, 2015) and incorporated Bayesian shrinkage methods to estimate unknown parameters simultaneously and to select important proteomic biomarkers of rapid disease progression. We identified a set of relevant clinical or demographic predictors and a proteomic biomarker for rapid decline of lung function. These biomarkers are moving forward past preclinical testing. Although these remarks stem from CF research, they demonstrate the potential for immediate applications and extensions of the authors' contribution to the analysis of medical monitoring data.

Kristin Kirchner (*Delft University of Technology*)

Asar, Bolin, Diggle and Wallin are to be congratulated for an inspiring contribution in the field of statistical modelling, presenting a methodology which combines the remarkable flexibility of non-Gaussian processes with computational efficiency.

Indeed, the computational benefits of the stochastic partial differential equation approach in the Gaussian case with differential operators as in equation (5) for fractional exponents $(2\phi + 1)/4 = k/2$ with

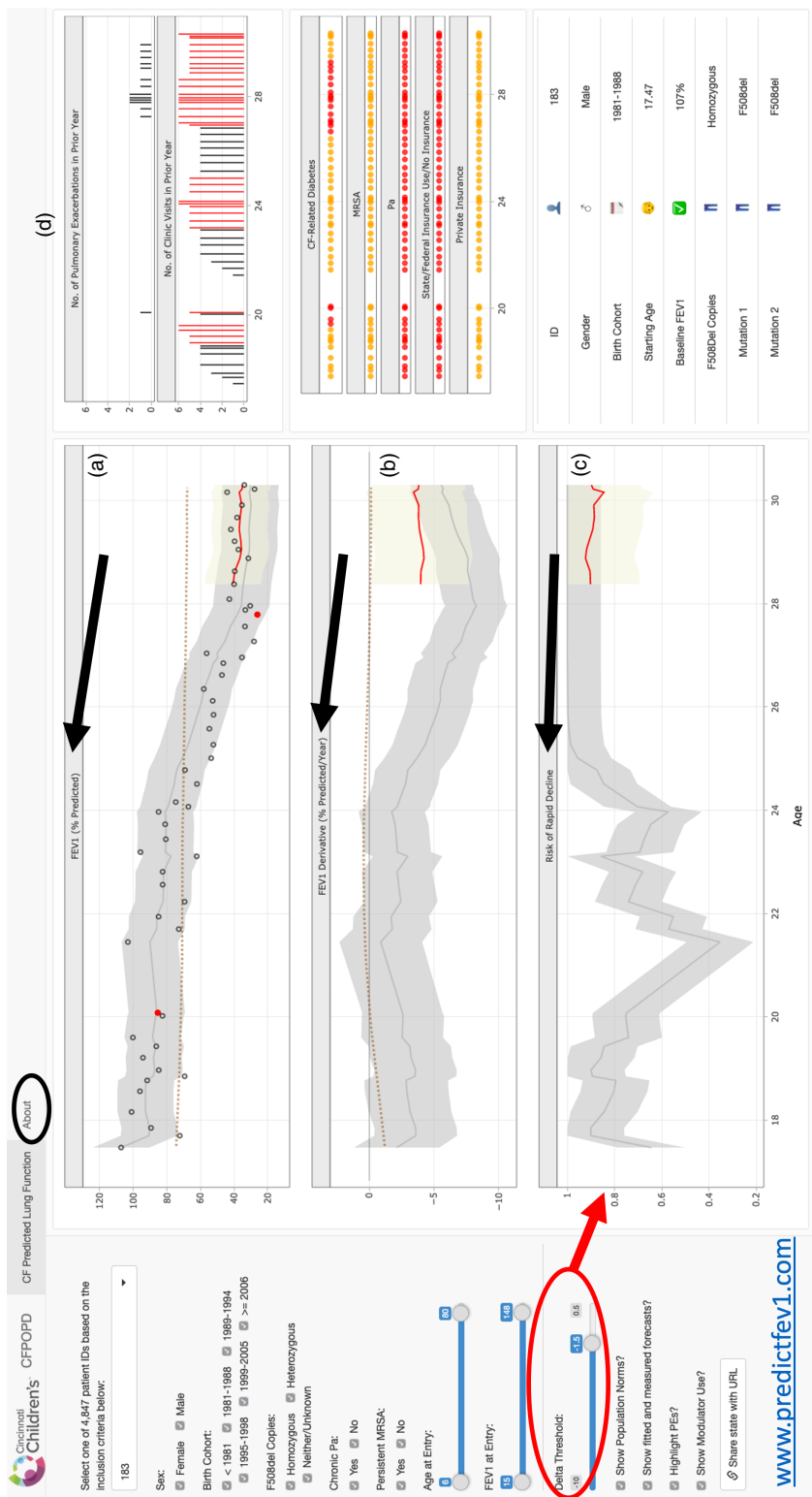


Fig. 12. R Shiny application (screen shot) displaying realtime risk of rapid decline in the individual CF patient (female case with genotype having no F508del alleles); the left-hand side provides user-specific selections for patients and demographic and clinical characteristics; the middle includes graphs corresponding to lung function trajectory, rate of change and predictive probability of rapid decline (threshold set to $\delta_P = -1.5\%$ predicted/year); the right-hand side shows a heat map of covariates included in the model for rolling numbers of exacerbations and clinic visits per year, infection with methicillin-resistant *Staphylococcus aureus*, use of state or federal insurance, diagnosis of CF-related diabetes and infection with *Pseudomonas aeruginosa*

$k \in \mathbb{N}$ have already been emphasized by Lindgren *et al.* (2011). Recent developments in numerical methods for fractional operators furthermore facilitate treating the whole admissible parameter range $\phi > 0$ while maintaining low computational complexity (Bolin and Kirchner, 2020; Bolin *et al.*, 2020; Herrmann *et al.*, 2020). Specifically, it is a well-known result that, in the non-fractional case $\mathcal{D}_1 = \kappa^2 - \partial^2/\partial t^2$, inverses of the *symmetric*, elliptic differential operator \mathcal{D}_1 , when discretized by means of a finite element method with continuous piecewise linear basis functions on the mesh $0 = s_1 < \dots < s_K = t_{\max}$ as used in equation (14), can be realized numerically with multilevel preconditioned iterative solvers in a complexity which is (essentially) linear in the number K of mesh nodes; see, for example Xu (1992). Recently, it has been shown that, given a prescribed accuracy, this complexity can also be achieved in the fractional case (Herrmann *et al.*, 2020).

The numerical inversion of the operator $\mathcal{D}_2 = \kappa + \partial/\partial t$ introduced in Section 3.2.1 and treated in a variational framework with a Petrov–Galerkin discretization in Appendix B, however, requires an additional discussion due to its *asymmetry*. The variational approach as described in Appendix B results from multiplying the ordinary differential equation

$$\dot{u}(t) + \kappa u(t) = f(t), \quad t \in (0, t_{\max}), \quad \text{subject to } u(0) = 0 \text{ (or } u(t_{\max}) = 0), \quad (32)$$

with a test function $v \in L^2(0, t_{\max})$, integrating over $(0, t_{\max})$, and employing the natural choice for a conforming Petrov–Galerkin discretization: continuous piecewise linear basis functions vanishing at 0 (or at t_{\max}) as the discrete trial space and piecewise constants as the discrete test space. It is worth noting that this variational approach is equivalent to the Crank–Nicolson scheme for the ordinary differential equation (32). In the non-fractional case $\mathcal{D}_2 = \kappa + \partial/\partial t$, an efficient numerical inversion of \mathcal{D}_2 can be realized with iterative solvers after

- (a) symmetrizing the problem and
- (b) appropriate preconditioning;

see Andreev (2014).

The matrix \mathbf{K} , whose inverse appears in the conditional distribution (15), results from discretizing the differential operator \mathcal{D}_1 or \mathcal{D}_2 . For the reasons above, multiplication with \mathbf{K}^{-1} , which is needed, for example, for likelihood evaluations or predictions, can be approximated efficiently.

To the best of my knowledge, numerical methods for fractional powers of the asymmetric operator $\kappa + \partial/\partial t$ are rare in the literature so far. This paper is a promising incentive to investigate statistical properties and computational algorithms for the fractional asymmetric model in future research.

Jorge Mateu (*University Jaume I, Castellón*)

Asar, Bolin, Diggle and Wallin are to be congratulated on a valuable contribution and thought-provoking paper on setting the pace for the extension of linear mixed effect models for non-Gaussian responses with repeated measures. They open this door by allowing any combination of the stochastic components of these linear models to be non-Gaussian. I would like to focus on the role of the stochastic process $W_i(t)$ and how the current framework can be adapted to other types of data and problems. The fact that W_i comes as a solution of a stochastic differential equation and known models, such as the integrated random-walk model or the Matérn model, can be obtained as particular cases in this framework brings direct connections to some extensions. One such extension is considering responses that are not continuous random variables, e.g. counting Poisson or negative binomial variables in combination with non-Gaussian random effects. Bivariate counts as responses define also a crucial extension, which is necessary in many real problems. But arguably the main extension could be to spatial processes, going from the stochastic real process $W_i(t)$ to a random field. In the same way that time ago techniques from longitudinal analysis were extended and adapted to the spatial framework (and I particularly refer to some work by Professor Diggle), the proposed (longitudinal) modelling framework can be extended to the spatial context (see Bolin and Wallin (2020)). It is easy to think that the individual-based stochastic process W_i could be living in the plane and defined by a particular random field. We can then formulate the model by using the stochastic partial differential equation representation of Gaussian Matérn fields. We have then room to develop spatially varying non-Gaussian repeated measures, and we can delineate some interesting details. For example, consider a non-stationary and/or anisotropic Gaussian random-field model that can model these characteristics on the response variable. By working on the properties of the underlying random field, we can obtain nice and extended versions of the models proposed. Another detail here is the concept of repeated measures in the response variable, which translates into the geostatistical context of the variance–nugget effect. And I finish by indicating that this would be another nice example of model-based geostatistics.

Giuseppe Storti and Pietro Coretto (*University of Salerno*)

We congratulate Asar, Bolin, Diggle and Wallin for their inspiring paper. Given the level of the paper we can add only a couple of remarks.

- (a) A deeper investigation of model selection could help to improve further the predictive ability of the model under analysis. For this, the preliminary estimates of the latent components, obtained from the basic linear Gaussian specification (eventually estimated without subsampling), might provide useful information for identifying the reference model among alternative distributional assumptions. In this respect, the exploratory analysis could be sharpened by considering the use of goodness-of-fit tests and information criteria. In addition, the complete-data Gaussian-based estimates of the model's components could be used to provide preliminary estimates of the generalized hyperbolic (GH) parameters that, in turn, could be employed to initialize the likelihood optimization and to speed convergence. Also, to alleviate the computational difficulties discussed in Appendix A.3, the optimization could be performed through a block-iterative procedure, where the GH parameters are separately estimated from the predicted components. For the normal-inverse Gaussian distribution, if computing time is an issue, the density parameters could be estimated by a method-of-moments estimator at the price of some restrictions on the admissible parameter space (Ghysels and Wang, 2014).
- (b) It is acknowledged that for some of these models (e.g. the GH specification) there are identifiability issues, which in practice translate into the common problem of multiple maxima of the log-likelihood function. This raises some questions about the *multiple-chain estimation* strategy that is central to stochastic gradient implementation. Starting from an initial guess Θ_0 , $\Theta_j^{(n)}$ and $\sigma^{(n)}$ are computed to define the algorithm's stopping criteria, standard errors and confidence intervals. The variance of Θ_j is approximated as $(\sigma_j^{(n)})^2 + \sigma_j^2$, i.e. the sum of the 'algorithmic' variance and the maximum likelihood (ML) variance based on the Fisher information matrix. Now suppose that Θ_0 is contained in a neighbourhood with barely separated multiple likelihood maxima. Under certain conditions, the ML estimation sequence may still be consistent for a local maximum, but the ML variance (the σ_j^2 -component) would be seriously affected unless Θ_0 is sufficiently close to one of the local maxima. The impression here is that the multiple-chain method increases the chance that, along the Monte Carlo runs, solutions might be attracted towards more than one of these local maxima. This mechanism may introduce some additional bias in the estimated uncertainty via $\sigma_j^{(n)}$, and this could happen even if Θ_0 is sufficiently close to one of the local maxima.

Zihao Wen and David L. Dowe (*Monash University, Clayton*)

We thank Asar, Bolin, Diggle and Wallin for their paper. Linear mixed effects models are commonly used for continuous repeated measurement data. The paper proposes a new method that allows the stochastic components to be decoupled and to come not just from a Gaussian distribution but also from a variety of non-Gaussian distributions. The authors also provide a stochastic gradient descent algorithm (Section 4.2) to implement maximum likelihood estimation. The method proposed seems to focus on a large data set (Section 4.2).

In Section 5, the authors validate and select models by visualizing Q - Q -plots. How would this method work if two plots are visually similar, and do the authors have a (proposed) quantitative way of automating this? Might the authors have any comment on the relative merits (for model selection) of instead using penalized likelihood approaches like the Akaike information criterion (Akaike, 1974), Bayesian information criterion (Schwarz, 1978) and perhaps also—using a Bayesian prior and (cf. Section 4.2) the Fisher information—minimum message length (Wallace and Freeman (1987), Wallace and Dowe (1999), section 6.1.2, Wallace and Dowe (2000), section 2, Fitzgibbon *et al.* (2004), Wallace (2005), chapter 5, Dowe *et al.* (2007), footnote 15, and Dowe (2011), section 6.3)?

Again in Section 5, it appears that model validation and selection are done by simulating Y_{ij}^S according to the model, and then comparing \hat{Z}_{ij} with \hat{Z}_{ij}^S . We would like to know the purported advantages of this method over choosing the maximum likelihood model.

Still in Section 5, when the authors say to simulate Y_{ij}^S according to the model, it appears that the simulation is based on (the training data set as) the complete data set. We would like to suggest that they might instead possibly do something of a cross-validation, dividing the data into a disjoint union of training data and test data, and then doing the simulation based on the test data set. Paraphrasing, we propose computing \hat{Z}_{ij} on the training data set, and then simulating data according to the model and values of the features in the test data set, and then computing \hat{Z}_{ij}^S on the simulated data. We would be curious to know what the authors think.

Regarding Table 7, some arguments for also including the logarithm of probability as a loss function—at least in the general case—are given in Dowe (2008), footnotes 175 and 176, Dowe (2011), section 3, and Dowe (2013), section 4.1, although the authors' results already seem quite clear.

Last, with current world circumstances, might the authors relate one of their models or a close variant to modelling Covid-19 coronavirus data?

The **authors** replied later, in writing, as follows.

We thank all the discussants for their interest in the paper and their helpful comments. We have arranged our reply under five broad topic headings, as follows.

Model interpretation

Among his many insightful comments, Daniel Farewell questions some aspects of our continuous time model formulation, including the existence (or not) of an instantaneous true state of nature and our formal separation of U_i and $W_i(t_{ij})$ in equation (2). From a strict mathematical perspective this separation is redundant, as is our separation of the two fixed effects terms $\mathbf{a}_i^T \boldsymbol{\alpha}$ and $\mathbf{d}_{ij}^T \boldsymbol{\beta}$ in the same equation. We gently disagree and, risking a charge of pedantry, mildly regret not initially writing this equation as

$$Y_{ij} = \{\mathbf{a}_i^T \boldsymbol{\alpha} + \mathbf{d}_i^T \mathbf{U}_i\} + \{\mathbf{x}_{ij}^T \boldsymbol{\beta} + W_i(t_{ij})\} + \sigma Z_{ij}$$

to emphasize the dual distinction between time constant and time-varying effects, and between observed (fixed) and unobserved (random) effects—the reason we did not do so is that this would have superficially excluded some widely used special cases, notably the random intercept and slope model of Laird and Ware (1982). If we admit the existence of measured variables in continuous time, we must surely admit the existence of unmeasured variables for which $W_i(t)$ acts as a general proxy.

However, there can be multiple interpretations of the random effects in combination with certain stochastic processes. For example, the sum of a Brownian motion $W_i(t)$, with initial value $W_i(0) = 0$, and a Gaussian random intercept U_i is equivalent to the same Brownian motion with initial value $W_i(0) = U_i$. Treating the U_i as the random intercept is incorrect since the time average of a subject realized Brownian motion will not be zero. An approach for handling such issues is to orthogonalize the random effects to obtain more easily interpretable results. One could for example add a sum-to-zero constraint on the Brownian motion to allow U_i to be interpreted as the mean. However, one should not do so with the fixed effects also, since this would incorrectly remove uncertainty (variance) of the estimate; see Sørbye *et al.* (2019) for a discussion about this in a spatial setting.

In the specific case of %FEV1, any time-varying effect must strictly be an integral over time of another stochastic process, implying a restriction on the legitimate class of covariance structures for $W_i(t)$. Whether this restriction matters in practice is less clear, provided that the focus of attention is on parametric inference rather than on prediction of unusual events, as was the case in our other application.

Similarly, the 'measurement error' term Z_{ij} is a misnomer, in the sense that in the model fitting process it captures both literal measurement error and (unidentifiable) stochastic variation on timescales that are shorter than the intervals between successive measurements. In a designed study, this ambiguity could be resolved by varying the follow-up intervals, including repeated measurements at effectively the same time where feasible, and conducting a frequency domain analysis of the residuals from a fixed-effect-only model. This could in turn inform the design of future studies to ensure that the follow-up protocol captures important components of the overall variation. In the absence of genuinely independent replication at a single time point, such as splitting a blood sample and running duplicate assays, the term 'measurement error' might more accurately be called 'high frequency variation'.

Other model formulations

As Jorge Mateu points out, a natural extension of the models proposed is to spatial data. In fact, the *ngme* package already supports models for both spatial data (as in Wallin and Bolin (2015)) and spatial multivariate data (as in Bolin and Wallin (2020)). Mateu also suggests investigating whether non-stationary Gaussian random fields could be constructed to capture the features of these spatial models. Because of the use of normal variance–mean mixtures in the construction, one could think of the non-Gaussian models in a Bayesian context as non-stationary Gaussian fields with a specific prior for spatially varying mean and variance function; see equation (15). This gives some intuition regarding the behaviour of the models but also suggests that it might be difficult to mimic the behaviour by using non-stationary Gaussian fields with smoothly varying parameters. Thus, non-stationary Gaussian models that could mimic the proposed

non-Gaussian models would probably need to have a large number of parameters that may be difficult to estimate.

A related point to the spatial extension is the comment from Neil Chada and Ajay Jasra regarding α -stable processes. These are very natural when it comes to stochastic partial differential equation representations given that the distribution of the random field is the same as that of the noise (since the distributions are closed under linear combinations). The main inferential issue is that, to our knowledge, the only distributions for which closed form expressions are available are the Gaussian and Cauchy families. This sets limits on parameter inference, on exploring posterior distributions and on latent variable formulations. It also explains why, in earlier work, we focused rather on closure under convolution to give two further distributional options (Bolin, 2014; Wallin and Bolin, 2015; Bolin and Wallin, 2020). For interpolation, the Cauchy distribution's heavy tails can accommodate clear jumps in the trajectories. However, for extrapolation the inferred jumps may become unrealistically large.

Arnošt Komárek suggests spline-based models as an alternative to our stochastic process formulation and argues that this will result in fewer parameters and easier interpretation for a practitioner. We respectfully disagree on both points. Regarding the parameters, a spline with K basis functions has K random effects. If these are Gaussian, then unless we make a sparse covariance structure assumption this generates $K(K+1)/2$ parameters, whereas in our formulations the random process has one or two parameters. Also, the representation of a spline depends on the spline basis as well as on the number and locations of the knots. The same could of course be said for our basis representation of the stochastic process. However, as the number of basis functions grows, the basis representation converges to a limiting process; hence there is no loss of interpretability. Regarding the interpretability for practitioners we argue to the contrary that with sufficient data we can estimate properties of our stochastic process models, such as their differentiability, that help with their interpretability, whereas for splines these properties are tied to the spline basis, and hence not determined by the data.

Measurement error in longitudinal covariates

Kuldeep Kumar raises this very relevant question. Suppose that a stochastic covariate $X(t)$ is being used to predict a stochastic outcome $Y(t)$. Then, it is legitimate to condition on the observed value of $X(t)$ and to make predictions on this basis, irrespective of the presence or absence of measurement error in $X(t)$. However, it is well known that, if $X(t) = X^*(t) + Z^*(t)$ where $X^*(t)$ is the error-free version of $X(t)$, then ignoring the measurement error term $Z^*(t)$ leads to biased estimation of the association between $X^*(t)$ and $Y(t)$. Within our general approach, the correct way to deal with this is to use a bivariate stochastic process model for $X(t)$ and $Y(t)$: see, for example, Bolin and Wallin (2020). However, the specific model formulation should depend on the scientific status of $X(t)$. If $X(t)$ and $Y(t)$ are coevolving biomarkers of equal interest, a symmetric formulation might be reasonable. In contrast, if there is a causal pathway from X to the outcome of interest, Y , a more natural formulation, using the square bracket notation to mean 'the distribution of', might be of the form $[X^*][X|X^*][Y|X^*]$, with the likelihood following by integration with respect to the unobserved X^* . Also the association between X^* and Y might need to be time lagged. Suppose, for example, that in the renal example of Section 6.2 we had data on longitudinal measurements of blood pressure, $X(t)$. High blood pressure damages susceptible kidneys, which in turn leads to loss of kidney function. Hence, a candidate model might be a model that measures the effect of X^* on Y as a weighted integral:

$$\int_0^\infty X^*(t-u)w(u;\theta)du,$$

with $w(u;\theta)$ monotone decreasing in u .

Computation

The numerical methods presented can probably be improved in various ways. We agree with Kristin Kirchner that incorporating state of the art numerical methods to sample the processes would reduce computation times, in particular for the spatial models that are implemented in the *ngme* package. We also encourage further investigations of numerical methods for models with fractional powers of asymmetric differential operators, which would be of direct use in our setting.

Regarding the use of stochastic optimization, we implemented several bespoke procedures because many of the standard procedures that we tested worked poorly for our applications. We performed initial tests with methods like Adam but ended up not using them because of stability problems. This may, however, be due to suboptimal implementations on our part, and we agree with Alex Stringer, Patrick Brown and

Jamie Stafford that a more systematic comparison of different stochastic optimization methods for our, and other, statistical models would be very valuable.

Running parallel optimizations enables the construction of suitable stopping criteria. Also, multiple chains could be used to construct preconditioners that do not introduce any bias. As is well known, the use of the same samples for estimating the preconditioner and the gradient can introduce bias in the estimates. Running multiple chains we could avoid this, e.g. by using the preconditioner computed in the first chain when updating the value of the second chain and vice versa. Running multiple parallel chains does not increase the computation time by much (given enough cores for the computations) but does increase memory requirements. This is, however, a smaller concern for us than for typical machine learning problems involving much larger data sets.

Model selection strategies

Giuseppe Storti and Pietro Coretto suggest the practical approach of first fitting the Gaussian model to the whole data and inspecting the distributions of the predicted values of U , W and Z . This is in contrast with our approach, where we simulate data under a fitted model (with subsampling for the renal application and without subsampling for the cystic fibrosis (CF) example) and inspect the QQ -plots of predicted values based on observed and simulated data sets. Envelopes of QQ -plots would help to understand whether apparent departure from the assumed model was due to chance.

Zihao Wen and David Dowe argue that different distributional assumptions could generate similar QQ -plots. Model selection criteria could be used for such scenarios, although there is always a limit to the ability of empirical procedures to discriminate between models. Also, model selection criteria measure the overall fit of an assumed model, whereas QQ -plots help us to assess the fit of individual terms. We note with interest their suggestion to apply cross-validation to our QQ -plot approach by simulating data for the validation set.

Giuseppe Storti and Pietro Coretto also raise the issue that the likelihood is not guaranteed to have a unique maximum. The use of multiple chains could help to detect the existence of multiple maxima and act as a warning that standard likelihood-based inferences may be unreliable. It is difficult to choose between different local maxima since we do not have access to the likelihood values. One approach for avoiding multiple maxima is to add a regularization term. As pointed out by Kumar, this also offers an approach to high dimensional variable selection.

Marked point process formulation, handling missing data

Daniel Farewell and Christine Chai both mention the issue of missing data, i.e. unrecorded measurements at prescheduled time points. In designed studies, missing values are self-evident. In observational studies the notion of a missing datum applies only if the data record that a measurement was made but not its value. Missing values are particularly problematic if they are *informative*, meaning that *conditional on all observed data* the fact that a measurement is missing and its (consequently unobserved) value are stochastically dependent. The literature on methods for dealing with missing values in longitudinal data is extensive; see, for example, Little (1995), Little and Rubin (2002) and Daniels and Hogan (2008).

Farewell also raised the related issue of informative follow-up times. In all our analyses, we assume that the timings of the repeated measures are either fixed by design, or otherwise stochastically independent of the measurement process. For observational data sets, e.g. data from electronic health records, this assumption might not hold. The behaviour of the renal data set that we analysed in our paper is suggestive of informative follow-up. For example, we see clusters of observations taken quite frequently: in some cases daily. The reason for this is not recorded in our data, but is quite likely to be related to a patient's current health status. Recognition of serial correlation in the measurement process alleviates, but not does not entirely remove, the potentially adverse effects of this on the accuracy of our inferences. A possible model-based solution is to couple a measurement model with an intensity submodel, in other words a marked point process where the points are the follow-up times and the marks are the repeated measurements. The analogous problem in a spatial setting has been termed *preferential sampling* (Diggle *et al.*, 2010). Contributions in a longitudinal setting include Lipsitz *et al.* (2002) and Lin *et al.* (2004).

Applications

We are pleased that Anna Choi, Tze Leung Lai and Zihao Wen are finding our methods useful in their study of Covid-19 transmission. We look forward to publication of their results in due course and would welcome their, or any other users', suggestions on possible modifications to the `ngme` package.

Emrah Geili and Rhonda Szczesniak point out that the concept of individual and time-specific prediction is important for the CF application. As CF is a rare disease, patients with CF are followed quite

closely; hence data quality is high and a large amount of data is available for each patient. For data sets with many repeated measures per patient, a time invariant random-effects assumption is unlikely to be suitable and the option to include a stochastic process component is, in our opinion, essential. In our application to the Danish CF data, we gave priority to estimation rather than prediction. We selected a model with normal inverse Gaussian distributions for all three stochastic terms. A model of this kind could provide an interesting extension to the Cincinnati Hospital case-study, and to the functionality of Geili and Szczesniak's very nice user interface.

References in the discussion

- Akaike, H. (1974) A new look at the statistical model identification. *IEEE Trans. Autom. Control*, **19**, 716–723.
- Andreev, R. (2014) Space-time discretization of the heat equation. *Numer. Alg.*, **67**, 713–731.
- Bolin, D. (2014) Spatial Matérn fields driven by non-Gaussian noise. *Scand. J. Statist.*, **41**, 557–579.
- Bolin, D. and Kirchner, K. (2020) The rational SPDE approach for Gaussian random fields with general smoothness. *J. Computat. Graph. Statist.*, **29**, 274–285.
- Bolin, D., Kirchner, K. and Kovács, M. (2020) Numerical solution of fractional elliptic stochastic PDEs with spatial white noise. *IMA J. Numer. Anal.*, **40**, 1051–1073.
- Bolin, D. and Wallin, J. (2020) Multivariate type G Matérn stochastic partial differential equation random fields. *J. R. Statist. Soc. B*, **82**, 215–239.
- Box, G. E. P. (1976) Science and statistics. *J. Am. Statist. Ass.*, **71**, 791–799.
- Chada, N. K., Lasanen, S. and Roininen, L. (2019) Posterior convergence analysis of α -stable sheet. *Preprint arXiv:1907.03086*.
- Daniels, M. J. and Hogan, J. W. (2008) *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Boca Raton: Chapman and Hall–CRC.
- Diggle, P. J., Menezes, R. and Su, T.-L. (2010) Geostatistical analysis under preferential sampling (with discussion). *Appl. Statist.*, **59**, 191–232.
- Diggle, P. J., Sousa, I. and Asar, O. (2015) Real-time monitoring of progression towards renal failure in primary care patients. *Biostatistics*, **16**, 522–536.
- Dowe, D. L. (2008) Foreword re C. S. Wallace. *Comput. J.*, **5**, 523–560.
- Dowe, D. L. (2011) MML, hybrid Bayesian network graphical models, statistical consistency, invariance and uniqueness. In *Philosophy of Statistics* (eds P. S. Bandyopadhyay and M. R. Forster), pp. 901–982. New York: Elsevier.
- Dowe, D. L. (2013) Introduction to Ray Solomonoff 85th memorial conference. In *Algorithmic Probability and Friends, Bayesian Prediction and Artificial Intelligence*, pp. 1–36. New York: Springer.
- Dowe, D. L., Gardner, S. and Oppy, G. R. (2007) Bayes not bust!: Why simplicity is no problem for Bayesians. *Br. J. Phil. Sci.*, **58**, 709–754.
- Fan, Y. and Li, R. (2012) Variable selection in linear mixed effects models. *Ann. Statist.*, **40**, 2043–2068.
- Farewell, D. M., Huang, C. and Didelez, V. (2017) Ignorability for general longitudinal data. *Biometrika*, **104**, 317–326.
- Fitzgibbon, L. J., Dowe, D. L. and Vahid, F. (2004) Minimum message length autoregressive model order selection. In *Proc. Int. Conf. Intelligent Sensing and Information Processing*, pp. 439–444. New York: Institute of Electrical and Electronics Engineers.
- Ghysels, E. and Wang, F. (2014) Moment-implied densities: properties and applications. *J. Bus. Econ. Statist.*, **32**, 88–111.
- Herrmann, L., Kirchner, K. and Schwab, C. (2020) Multilevel approximation of Gaussian random fields: fast simulation. *Math. Modls Meth. Appl. Sci.*, **30**, 181–223.
- Jacobsen, M. (2006) Probability and its applications. In *Point Process Theory and Applications*. Boston: Birkhäuser.
- Kingma, D. P. and Ba, J. (2015) Adam: a method for stochastic optimization. *Int. Conf. Learning Representations*.
- Laird, N. M. and Ware, J. H. (1982) Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.
- Lin, H., Scharfstein, D. O. and Rosenheck, R. A. (2004) Analysis of longitudinal data with irregular, outcome-dependent follow-up. *J. R. Statist. Soc. B*, **66**, 791–813.
- Lindgren, F., Rue, H. and Lindström, J. (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach (with discussion). *J. R. Statist. Soc. B*, **73**, 423–498.
- Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J. G., Gelber, R. and Lipshultz, S. (2002) Parameter estimation in longitudinal studies with outcome-dependent follow-up. *Biometrics*, **58**, 621–630.
- Little, R. J. A. (1995) Modelling the drop-out mechanism in repeated-measures studies. *J. Am. Statist. Ass.*, **90**, 1112–1121.
- Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis with Missing Data*, 2nd edn. New York: Wiley.
- Pan, J. and Mackenzie, G. (2003) On modelling mean-covariance structures in longitudinal studies. *Biometrika*, **90**, 239–244.

- Richardson, T. S. and Robins, J. M. (2013) Single world intervention graphs (SWIGs): a unification of the counterfactual and graphical approaches to causality. *Working Paper 128*. Center for the Statistics and the Social Sciences, University of Washington, Seattle.
- Roininen, L., Girolami, M., Lasanen, S. and Markkanen, M. (2019) Hyperpriors for Matérn fields with applications in Bayesian inversion. *Inverse Problems and Imaging*.
- Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.
- Sørbye, S. H., Illian, J. B., Simpson, D. P., Burslem, D. and Rue, H. (2019) Careful prior specification avoids incautious inference for log-Gaussian Cox point processes. *Appl. Statist.*, **68**, 543–564.
- Stuart, A. M. (2010) Inverse problems: a Bayesian perspective. *Acta Numer.*, **19**, 451–559.
- Szczesniak, R. D., Su, W., Brokamp, C., Keogh, R. H., Pestian, J. P., Seid, M., Diggle, P. J. and Clancy, J. P. (2019) Dynamic predictive probabilities to monitor rapid cystic fibrosis disease progression. *Statist. Med.*, to be published.
- Verbeke, G. and Lesaffre, E. (1997) The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computnl Statist. Data Anal.*, **23**, 541–556.
- Wallace, C. S. (2005) *Statistical and Inductive Inference by Minimum Message Length*. New York: Springer Science and Business Media.
- Wallace, C. S. and Dowe, D. L. (1999) Minimum message length and Kolmogorov complexity. *Comput. J.*, **42**, 270–283.
- Wallace, C. S. and Dowe, D. L. (2000) MML clustering of multistate, Poisson, von Mises circular and Gaussian distributions. *Statist. Comput.*, **10**, 73–83.
- Wallace, C. S. and Freeman, P. R. (1987) Estimation and inference by compact coding. *J. R. Statist. Soc. B*, **49**, 240–252.
- Xu, J. (1992) Iterative methods by space decomposition and subspace correction. *SIAM Rev.*, **34**, 581–613.