covBM: incorporating Brownian motion components into 'nlme' models

Oliver Stirrup
MRC Clinical Trials Unit at UCL, University College London, London, UK

1 Introduction

Longitudinal data are now widely analysed using linear mixed models, with 'random slopes' models particularly common. These models can successfully account for the dependency that arises when repeated observations are made over time on each individual in a dataset, but make strong assumptions regarding the nature of this dependency. In the context of modelling CD4 cell counts over time in human immunodeficiency virus (HIV)-positive patients, it has been shown that the incorporation of non-stationary stochastic processes such as Brownian motion or integrated Ornstein–Uhlenbeck (IOU) processes into the modelling framework can lead to a very substantial improvement in model fit [1;2]. Recently, the use of a fractional Brownian motion component has been shown to provide a further improvement [3]. However, these extensions to the standard linear mixed model have not been widely used in practice, and are not readily implemented in current statistical software programs. The presence of such a component in a model for longitudinal data implies that the progress of the state of the underlying biological system for each individual does not follow a deterministic relationship with time, but rather follows an unpredictable stochastic path.

The `nlme` package [4] for R allows the user to fit a wide range of linear and non-linear mixed effects models, with in-depth documentation and a wealth of examples provided in the accompanying book by Pinheiro and Bates [5]. As well as incorporating within-subject dependence resulting from the inclusion of 'random effects' in a specified model, `nlme` also allows for a correlation structure to be imposed on the residual error terms (with estimation of any associated parameters) and for the residual error variance to be modelled as a function of variables in the data under consideration. It is even possible for the user to create their own correlation structures or variance functions for inclusion in the estimation of models in `nlme`.

It is possible to implement user-defined correlation structures in `nlme` to obtain point estimates of the parameters in linear and non-linear mixed effects models incorporating Brownian motion or IOU processes. However, some further additions to the original `nlme` code are required to obtain confidence intervals for the natural model parameters and to return a fitted model object that reports the natural parameters upon use of `print` or `summary`. The `covBM` package provides wrappers for the standard `nlme` functions in order to achieve these goals.

In Section 2, the characteristics of the statistical models under consideration are specified, and in Section 3, examples are provided to illustrate use of the functions provided in `covBM` to fit such models.

2 Model description

2.1 Scaled Brownian motion

Brownian motion (also known as a Wiener process) is a non-stationary stochastic process that constitutes a continuous-time generalisation of a simple random walk [6], in which successive increments are independent of the history of the process. When considered in terms of a given set of

observation points, a scaled Brownian motion process, denoted W_t at time t, is defined by the properties:

$$\begin{aligned} W_0 &= 0 \\ W_t - W_s &\sim N(0, \kappa(t-s)) \text{ for } 0 \leq s < t. \end{aligned}$$

The process starts at zero at time (t) zero, and increments of the process are stationary, independent (for disjoint periods of time) and normally distributed with mean zero and variance equal to the difference in time between observation points scaled by a constant factor κ. These conditions lead to the following characteristics:

$$\begin{aligned} \mathrm{E}[W_t] &= 0 \\ \mathrm{Var}[W_t] &= \kappa t \\ \mathrm{Cov}[W_s, W_t] &= \kappa * \min(s,t). \end{aligned}$$

The distribution of a set of n observations relating to a given series of time points therefore follows a multivariate normal distribution with a mean vector of n zeros and covariance matrix defined by the formulae given above.

2.2 Scaled fractional Brownian motion

Fractional Brownian motion represents a generalisation of a Brownian motion process in which increments for disjoint time periods are not constrained to be independent, although they do remain stationary. The process was introduced by Mandelbrot and van Ness [7]. The characteristics of a fractional Brownian motion process are determined by an additional parameter, referred to as H or 'the Hurst index', that may take a value in the range (0,1). Standard Brownian motion represents a special case of fractional Brownian motion, corresponding to $H = \frac{1}{2}$. As for standard Brownian motion, the expectation of the value of the process is zero for all points in time.

When $H < \frac{1}{2}$, successive increments of the process are negatively correlated. This has the consequence, firstly, that the path of the trajectory appears 'jagged' and, secondly, that realisations of the process tend to revert towards the mean of zero. For $H > \frac{1}{2}$, successive increments of the process are positively correlated. This means that the path of the process has a relatively 'smooth' appearance, and also that realisations of the process tend to diverge away from zero.

As for Brownian motion, a scale parameter (κ) can be added to the standard definition of fractional Brownian motion, corresponding to the variance of the process at $t = 1$. We may then characterise the properties of the process as follows:

$$\begin{aligned} W_0 &= 0 \\ \mathrm{E}[W_t] &= 0 \\ \mathrm{Var}[W_t] &= \kappa \left|t\right|^{2H} \\ \mathrm{Cov}[W_s, W_t] &= \frac{\kappa}{2}\left(\left|s\right|^{2H} + \left|t\right|^{2H} - \left|t-s\right|^{2H}\right). \end{aligned}$$

2.3 Integrated Ornstein–Uhlenbeck process

The IOU process is another non-stationary Gaussian stochastic process that has also been used to model CD4 counts in HIV-positive patients, a full description is provided by Taylor *et al.* [1]. The process has the following characteristics:

$$\begin{aligned} W_0 &= 0 \\ \mathrm{E}[W_t] &= 0 \\ \mathrm{Var}[W_t] &= \frac{\kappa}{\alpha^3}\left(\alpha t + e^{-\alpha t} - 1\right) \\ \mathrm{Cov}[W_s, W_t] &= \frac{\kappa}{2\alpha^3}\left(2\alpha * \min(s,t) + e^{-\alpha t} + e^{-\alpha s} - 1 - e^{-\alpha|t-s|}\right). \end{aligned}$$

We have used the symbol κ to denote the variance scaling parameter (σ^2 was used by Taylor *et al.*[1]). The α parameter determines the extent to which the process reverts towards its mean value. For values of α approaching infinity, the process is equivalent to scaled Brownian motion, whereas for values of α approaching zero the process is equivalent to a random slopes model (without a random intercept)[1].

2.4 Marginal distribution

For models incorporating Gaussian processes such as Brownian motion, the fact that the marginal distribution of the full vector of observations of the outcome variable is multivariate normal (MVN) means that parameter estimation can be achieved through adjustment of the methods used for standard linear mixed models. The linear mixed model for longitudinal data can be expressed in the form[8]:

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i \\ \mathbf{b}_i &\sim MVN(\mathbf{0},\, \boldsymbol{\Psi}) \\ \mathbf{e}_i &\sim MVN(\mathbf{0},\, \mathbf{R}_i). \end{aligned} \tag{1}$$

Here, $\mathbf{y}_i$ represents the vector of n_i observations for the i^{th} individual, $\mathbf{X}_i$ represents their design matrix for the 'fixed effects' parameters $\boldsymbol{\beta}$, $\mathbf{Z}_i$ represents the subset of the columns of the design matrix associated with the 'random effects' for each individual $\mathbf{b}_i$ and $\mathbf{e}_i$ is the vector of residual errors for each measurement occasion. The vectors of random effects $\mathbf{b}_1, \mathbf{b}_2 ... \mathbf{b}_N$ and residual errors $\mathbf{e}_1, \mathbf{e}_2 ... \mathbf{e}_N$ for each of the N individuals are independent of one another. It can be easily shown that this formulation leads to the following marginal distribution for $\mathbf{y}_i$:

$$\mathbf{y}_i \sim MVN(\mathbf{X}_i\boldsymbol{\beta},\, \mathbf{Z}_i\boldsymbol{\Psi}\mathbf{Z}_i^{\mathrm{T}} + \mathbf{R}_i).$$

When linear mixed models are fitted to longitudinal data, it is common to assume that the residual errors for each observation within each individual, $\mathbf{e}_i$, are independent and with constant variance, σ^2, i.e. $\mathbf{R}_i$ as defined in (1) is equal to $\sigma^2\mathbf{I}_{n_i}$. However, other forms for $\mathbf{R}_i$ are widely used, particularly for the analysis of longitudinal or spatial data, for example the exponential correlation structure[5].

The remaining variability in the model, once the random effects have been accounted for, can also be subdivided into a component relating to a Gaussian process (independent of other model components) with expectation zero for all time points and an independent residual error for each observation. Defining $\boldsymbol{\Sigma_i}$ as the covariance matrix resulting from the chosen Gaussian process and set of time points $\mathbf{t}_i$ for the i^{th} individual, the linear mixed model can then be expressed as:

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + W_i[\mathbf{t}_i] + \mathbf{e}_i \\ \mathbf{b}_i &\sim MVN(\mathbf{0},\, \boldsymbol{\Psi}) \\ W_i[\boldsymbol{t}_i] &\sim MVN(\mathbf{0},\, \boldsymbol{\Sigma}_i) \\ \mathbf{e}_i &\sim MVN(\mathbf{0},\, \sigma^2\mathbf{I}_{n_i}), \end{aligned} \tag{2}$$

with marginal distribution:

$$\mathbf{y}_i \sim MVN(\mathbf{X}_i\boldsymbol{\beta},\, \mathbf{Z}_i\boldsymbol{\Psi}\mathbf{Z}_i^{\mathrm{T}} + \boldsymbol{\Sigma}_i + \sigma^2\mathbf{I}_{n_i}).$$

Although here we have focused on the marginal distribution for linear mixed models that incorporate a stochastic process, similar adjustment of the multivariate normal residual error distribution (i.e. $\mathbf{R}_i$) can also be made for non-linear mixed effects models.

3 Examples

3.1 lmeBM function

The `lmeBM` function is a wrapper for the `lme.formula` function from the `nlme` package, i.e. the `lme` function as used with a formula argument to specify the desired model; and the various

arguments can be used in exactly the same way as the original `nlme` function. However, `lmeBM` allows Brownian motion, fractional Brownian motion or IOU process components to be added to a model.

Included in the `covBM` package is a dataset of serial CD4 counts obtained in HIV-positive children. This dataset is discussed in *Data Analysis Using Regression and Multilevel/Hierarchical Models* by Andrew Gelman and Jennifer Hill[9], and the original is available online from the home page of this book. In the present package, rows with missing values of 'CD4CNT' (CD4 count on original scale), 'visage' (age of child in years at given visit) or 'baseage' (age of child in years at initial visit) have been removed.

```
> library(covBM)
> head(cd4)
```

```
  newpid   visage treatmnt CD4CNT  baseage   sqrtcd4         t
1      1 5.330833        1    626 3.910000 25.019992 1.4208333
2      1 5.848333        1    220 3.910000 14.832397 1.9383333
3      2 3.565000        2     30 3.565000  5.477226 0.0000000
4      2 3.778333        2      4 3.565000  2.000000 0.2133333
5      3 6.124167        1    714 6.124167 26.720778 0.0000000
6      3 6.354167        1    523 6.124167 22.869193 0.2300000
```

We will consider models for square root-transformed CD4 counts 'sqrtcd4', as this provides a better approximation to the normal distribution, in terms of the time elapsed in years since the initial visit 't'. The variable 'newpid' provides unique patient identifiers. The 'treatmnt' variable indicates whether that child was a control (==1) or given a zinc supplment (==2). However, this variable is not considered below.

First, we fit a standard 'random slopes' linear mixed model, using the `lme` function from the `nlme` package. We choose here to obtain the maximum likelihood parameter estimates throughout, although restricted maximum likelihood estimation could also be implemented using the argument `method=="REML"`.

```
> RS_model<-lme(sqrtcd4~t, data=cd4, random=~t|newpid, method="ML")
> RS_model
```

```
Linear mixed-effects model fit by maximum likelihood
  Data: cd4
  Log-likelihood: -3424.766
  Fixed: sqrtcd4 ~ t
(Intercept)           t
  30.664754   -5.556963

Random effects:
 Formula: ~t | newpid
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 12.606196 (Intr)
t            5.792515 -0.375
Residual     5.354337

Number of Observations: 976
Number of Groups: 226
```

We then fit a 'random slopes' linear mixed model with additional inclusion of a scaled Brownian motion process. This requires the `covariance=covBM` argument using the `lmeBM` function, which exactly follows the `lme` syntax. The parameter estimates for the model do not converge when using the default optimiser in this dataset, but the model can be successfully fitted using the `control=list(opt="optim")` argument.

```
> BM_model<-lmeBM(sqrtcd4~t, data=cd4, random=~t|newpid,
+                 covariance=covBM(form=~t|newpid), method="ML",
+                 control=list(opt="optim"))
> BM_model

Linear mixed-effects model fit by maximum likelihood
  Data: cd4
  Log-likelihood: -3421.276
  Fixed: sqrtcd4 ~ t
(Intercept)           t
  30.726746   -5.505073

Random effects:
 Formula: ~t | newpid
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 12.675137 (Intr)
t            3.362038 -0.732
Residual     4.850621

Stochastic process component: covBM
 Formula: ~t | newpid
 Parameter estimate(s):
   Kappa
34.92393
Number of Observations: 976
Number of Groups: 226
```

A further generalisation of the model to incorporate a fractional Brownian motion process can also be considered:

```
> fBM_model<-lmeBM(sqrtcd4~t, data=cd4, random=~t|newpid,
+                 covariance=covFracBM(form=~t|newpid), method="ML",
+                 control=list(opt="optim"))
> fBM_model

Linear mixed-effects model fit by maximum likelihood
  Data: cd4
  Log-likelihood: -3420.997
  Fixed: sqrtcd4 ~ t
(Intercept)           t
  30.763016   -5.479037

Random effects:
 Formula: ~t | newpid
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 12.727100 (Intr)
t            3.272245 -0.83
Residual     4.551875

Stochastic process component: covFracBM
 Formula: ~t | newpid
 Parameter estimate(s):
      Kappa Hurst index
```

```
 40.8411824   0.3776367
Number of Observations: 976
Number of Groups: 226
```

The fitted model objects created using the `lmeBM` function are of class `"lme"`, and so all the usual `nlme` Methods can be used to extract and view useful information. For example, `anova.lme` can be used to compare a set of fitted models:

```
> anova(RS_model, BM_model, fBM_model)

          Model df      AIC      BIC    logLik   Test  L.Ratio p-value
RS_model      1  6 6861.531 6890.832 -3424.766
BM_model      2  7 6856.552 6890.736 -3421.276 1 vs 2 6.979464  0.0082
fBM_model     3  8 6857.993 6897.061 -3420.997 2 vs 3 0.558621  0.4548
```

Both the likelihood ratio tests and a comparison of Akaike's information criterion (AIC) values suggest that the model including a Brownian motion process should be chosen above a standard random slopes model, but that there is not evidience to support the generalisation to a fractional Brownian motion process. This conclusion is also supported by inspection of the approximate 95 % confidence intervals of parameter estimates for the fractional Brownian motion model, as the confidence interval for the H-index is inclusive of 0.5 (the value for a standard Brownian motion process).

```
> intervals(fBM_model)$corStruct

                lower       est.      upper
Kappa       19.37188236 40.8411824 86.1042900
Hurst index  0.05989256  0.3776367  0.8524892
attr(,"label")
[1] "Correlation structure:"
```

The random slopes model incorporating an IOU process returns a high estimate of the α parameter, and does not show an improvement in fit relative to the scaled Brownian motion model.

```
> IOU_model<-lmeBM(sqrtcd4~t, data=cd4, random=~t|newpid,
+                 covariance=covIOU(form=~t|newpid), method="ML",
+                 control=list(opt="optim"))
> IOU_model

Linear mixed-effects model fit by maximum likelihood
  Data: cd4
  Log-likelihood: -3421.164
  Fixed: sqrtcd4 ~ t
(Intercept)           t
  30.721825   -5.490878

Random effects:
 Formula: ~t | newpid
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 12.655067 (Intr)
t            2.879292 -0.877
Residual     4.886538

Stochastic process component: covIOU
```

```
 Formula: ~t | newpid
 Parameter estimate(s):
      Kappa       Alpha
23758.24004    24.62638
Number of Observations: 976
Number of Groups: 226
```

```
> anova(BM_model, IOU_model)
```

```
          Model df      AIC      BIC    logLik   Test  L.Ratio p-value
BM_model      1  7 6856.552 6890.736 -3421.276
IOU_model     2  8 6858.327 6897.395 -3421.164 1 vs 2 0.224372  0.6357
```

3.2 nlmeBM function

The `nlmeBM` function is a wrapper for the `nlme.formula` function from the `nlme` package. As for `lmeBM`, `nlmeBM` allows Brownian motion or fractional Brownian motion components to be added to a non-linear mixed effects model.

As an illustrative example, we consider the `Milk` dataset available in the `nlme` package. This dataset is discussed in Chapter 5 of Diggle *et al.*[10], and contains measurements of the protein concentration of the milk of a number of cows assessed weekly following calving. The cows are divided into groups according to diet, but we ignore this for the sake of simplicity. We fit an asymptotic regression function, using `SSasmyp` from `nlme`, with three fixed effects parameters: `Asym` representing the horizontal asymptote for large values of the time variable, `R0` representing the response at time zero and `lrc` representing the natural logarithm of the rate constant (see Pinheiro and Bates[5] for further details). We consider an initial model with independent errors of constant variance and a second model with correlated errors following a continuous autoregressive process, both fit using the `nlme` function. Thirdly, we consider a model including a fractional Brownian motion process within each cow in addition to independent residual errors, using the `covariance=covFracBM` argument for `nlmeBM`. A subject-specific 'random effect' is assigned to the asymptote parameter in each of the models.

```
> Model_1<-nlme(protein ~ SSasymp(Time, Asym, R0, lrc), data=Milk,
+                                fixed = Asym + R0 + lrc ~ 1, random = Asym ~ 1|Cow,
+                            start = c(Asym = 3.5, R0 = 4, lrc = -1))
> Model_2<-nlme(protein ~ SSasymp(Time, Asym, R0, lrc), data=Milk,
+                                fixed = Asym + R0 + lrc ~ 1, random = Asym ~ 1|Cow,
+                                correlation=corCAR1(form=~Time|Cow),
+                            start = c(Asym = 3.5, R0 = 4, lrc = 0))
> Model_3<-nlmeBM(protein ~ SSasymp(Time, Asym, R0, lrc), data=Milk,
+                                fixed = Asym + R0 + lrc ~ 1, random = Asym ~ 1|Cow,
+                                covariance=covFracBM(form=~Time|Cow),
+                            start = c(Asym = 3.5, R0 = 4, lrc = -1))
> AIC(Model_1)

[1] 301.4711

> AIC(Model_2)

[1] -18.96245

> AIC(Model_3)

[1] -23.20265

> Model_3
```

```
Nonlinear mixed-effects model fit by maximum likelihood
  Model: protein ~ SSasymp(Time, Asym, R0, lrc)
  Data: Milk
  Log-likelihood: 18.60133
  Fixed: Asym + R0 + lrc ~ 1
      Asym         R0        lrc
3.34894647 4.72812704 0.03811161

Random effects:
 Formula: Asym ~ 1 | Cow
                Asym     Residual
StdDev: 6.779466e-08 5.609692e-05

Stochastic process component: covFracBM
 Formula: ~Time | Cow
 Parameter estimate(s):
      Kappa Hurst index
 0.07054058  0.16214425
Number of Observations: 1337
Number of Groups: 79
```

On the basis of the AIC values, the model including the fractional Brownian motion component provides the best fit to the data of those considered here.

References

[1] Taylor JMG, Cumberland WG, and Sy JP. A stochastic model for analysis of longitudinal AIDS data. *J Am Stat Assoc*, 89, 727–736 1994.

[2] Babiker AG, Emery S, Fätkenheuer G, Gordin FM, Grund B, Lundgren JD, Neaton JD, Pett SL, Phillips A, Touloumi G, and Vjechaj MJ; INSIGHT START Study Group. Considerations in the rationale, design and methods of the strategic timing of antiretroviral treatment (START) study. *Clin Trials*, 10 (1 Suppl):S5–S36, 2013.

[3] Stirrup OT, Babiker AG, Carpenter JR, and Copas AJ. Fractional brownian motion and multivariate-t models for longitudinal biomedical data, with application to cd4 counts in hiv-patients. *Statistics in Medicine*, page (in press), 2015.

[4] Pinheiro J, Bates D, DebRoy S, Sarkar D, and R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*, 2014. R package version 3.1-117.

[5] Pinheiro J and Bates D. *Mixed-Effects Models in S and S-PLUS*. Springer, 2000.

[6] Grimmett G and Stirzaker D. *Probability and Random Processes*, page 370. Oxford University Press, third edition, 2001.

[7] Mandelbrot B and van Ness JW. Fractional brownian motions, fractional noises and applications. *SIAM Review*, 10:422–437, 1968.

[8] Laird NM and Ware JH. Random-effects models for longitudinal data. *Biometrics*, 38:963–974, 1982.

[9] Gelman A and Hill J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Home page: http://www.stat.columbia.edu/∼gelman/arm/. Cambridge University Press, 2006.

[10] Diggle PJ, Heagerty P, Liang K-Y, and Zeger SL. *Analysis of Longitudinal Data*. Oxford University Press, second edition, 2002.