Linear mixed model implementation in lme4

Douglas Bates
Department of Statistics
University of Wisconsin – Madison

October 20, 2009

Abstract

We describe the form of the linear mixed-effects and generalized linear mixed-effects models fit by `lmer` and give details of the representation and the computational techniques used to fit such models. These techniques are illustrated on several examples.

```
Matrix_NS: <environment: namespace:Matrix>
```

1 A simple example

The `Rail` data set from the MEMSS package is described in Pinheiro and Bates (2000) as consisting of three measurements of the travel time of a type of sound wave on each of six sample railroad rails. We can examine the structure of these data with the `str` function

```
> str(Rail)
'data.frame':        18 obs. of  2 variables:
 $ Rail  : Factor w/ 6 levels "A","B","C","D",..: 1 1 1 2 2 2 3 3 3 4 ...
 $ travel: num  55 53 54 26 37 32 78 91 85 92 ...
```

Because there are only three observations on each of the rails a dotplot (Figure 1) shows the structure of the data well.

```
> print(dotplot(reorder(Rail, travel) ~ travel, Rail, xlab = "Travel time (ms)",
+     ylab = "Rail"))
```

In building a model for these data

```
> Rail
```

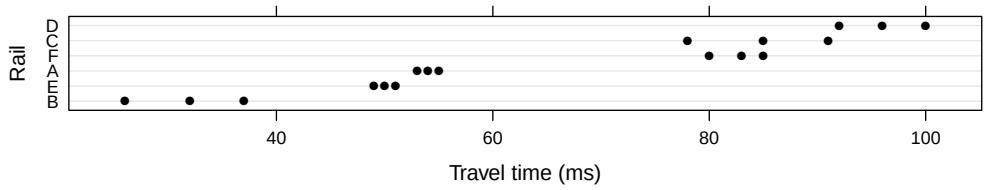

Figure 1: Travel time of sound waves in a sample of six railroad rails. There were three measurements of the travel time on each rail. The order of the rails is by increasing mean travel time.

	Rail	travel
1	A	55
2	A	53
3	A	54
4	B	26
5	B	37
6	B	32
7	C	78
8	C	91
9	C	85
10	D	92
11	D	100
12	D	96
13	E	49
14	E	51
15	E	50
16	F	80
17	F	85
18	F	83

we wish to characterize a typical travel time, say μ, for the population of such railroad rails and the deviations, say $b_i, i = 1, \ldots, 6$ of the individual rails from this population mean. Because these specific rails are not of interest by themselves as much as the variation in the population we model the b_i, which are called the "random effects" for the rails, as having a normal (also called "Gaussian") distribution of the form $\mathcal{N}(0, \sigma_b^2)$. The jth measurement on the ith rail is expressed as

$$y_{ij} = \mu + b_i + \epsilon_{ij} \quad b_i \sim \mathcal{N}(0, \sigma_b^2), \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \quad i = 1, \ldots, 6 \; j = 1, \ldots, 3 \tag{1}$$

The parameters of this model are μ, σ_b^2 and σ^2. Technically the $b_i, i = 1, \ldots, 6$ are not parameters but instead are considered to be unobserved random variables for which we form "predictions" instead of "estimates".

To express generalizations of models like (1) more conveniently we switch to a matrix/vector representation in which the 18 observations of the travel time form the response vector $\boldsymbol{y}$, the fixed-effect parameter μ forms a 1-dimensional column vector $\boldsymbol{\beta}$ and the six random effects $b_i, i = 1, \ldots, 6$ form the random effects vector $\boldsymbol{b}$. The structure of the data and the values of any covariates (none are used in this model) are used to create model matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$.

Using these vectors and matrices and the 18-dimensional vector $\boldsymbol{\epsilon}$ that represents the per-observation noise terms the model becomes

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}, \left(\mathbf{0}, \sigma^2 \boldsymbol{I}\right), \ \boldsymbol{b} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \boldsymbol{\Sigma}\right) \text{ and } \boldsymbol{b} \perp \boldsymbol{\epsilon} \qquad (2)$$

In the general form we write p for the dimension of $\boldsymbol{\beta}$, the fixed-effects parameter vector, and q for the dimension of $\boldsymbol{b}$, the vector of random effects. Thus the model matrix $\boldsymbol{X}$ has dimension $n \times p$, the model matrix $\boldsymbol{Z}$ has dimension $n \times q$ and the relative variance-covariance matrix, $\boldsymbol{\Sigma}$, for the random-effects has dimension $q \times q$. The symbol $\perp$ indicates independence of random variables and $\mathcal{N}$ denotes the multivariate normal (Gaussian) distribution.

We say that matrix $\boldsymbol{\Sigma}$ is the relative variance-covariance matrix of the random effects in the sense that it is the variance of $\boldsymbol{b}$ relative to σ^2, the scalar variance of the per-observation noise term $\boldsymbol{\epsilon}$. Although it size, q, can be very large, $\boldsymbol{\Sigma}$ is highly structured. It is symmetric, positive semi-definite and zero except for the diagonal elements and certain elements close to the diagonal.

1.1 Fitting the model and examining the results

The maximum likelihood estimates for parameters in model (1) fit to the `Rail` data are obtained as

```
> RmlML <- lmer(travel ~ 1 + (1 | Rail), Rail, REML = FALSE, verbose = TRUE)
  0:     149.28908: 0.942809
  1:     137.53112:  1.94281
  2:     132.38870:  2.85077
  3:     129.94249:  3.73815
  4:     128.94483:  4.52610
  5:     128.62895:  5.12722
  6:     128.56577:  5.47713
  7:     128.56016:  5.60451
  8:     128.56004:  5.62581
  9:     128.56004:  5.62686
 10:     128.56004:  5.62686
 11:     128.56004:  5.62686
 12:     128.56004:  5.62686
```

In this fit we have set `verbose = TRUE` indicating that information on the progress of the iterations should be printed after every iteration. Each line gives the iteration number, the value of the deviance (negative twice the log-likelihood) and the value of the parameter s which is the standard deviation of the random effects relative to the standard deviation of the residuals.

The printed form of the model

```
> Rm1ML

Linear mixed model fit by maximum likelihood
Formula: travel ~ 1 + (1 | Rail)
   Data: Rail
   AIC   BIC logLik deviance REMLdev
 134.6 137.2 -64.28    128.6   122.2
Random effects:
 Groups   Name        Variance Std.Dev.
 Rail     (Intercept) 511.861  22.6243
 Residual              16.167   4.0208
Number of obs: 18, groups: Rail, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)   66.500      9.285   7.162
```

provides additional information about the parameter estimates and some of the measures of the fit such as the log-likelihood (-64.28), the deviance for the maximum likelihood criterion (128.6), the deviance for the REML criterion (122.2), Akaike's Information Criterion (AIC= 132.6) and Schwartz's Bayesian Information Criterion (BIC= 134.3).

The transpose of the model matrix $\boldsymbol{Z}$ is stored as a sparse matrix in the `Zt` slot of the fitted model. For this model $\boldsymbol{Z}$ is simply the matrix of indicators of the levels of the `Rail`.

```
> Rm1ML@Zt

6 x 18 sparse Matrix of class "dgCMatrix"

[1,] 1 1 1 . . . . . . . . . . . . . . . . .
[2,] . . . 1 1 1 . . . . . . . . . . . . . .
[3,] . . . . . . 1 1 1 . . . . . . . . . . .
[4,] . . . . . . . . . 1 1 1 . . . . . . . .
[5,] . . . . . . . . . . . . 1 1 1 . . . . .
[6,] . . . . . . . . . . . . . . . . . 1 1 1

> as(Rail[["Rail"]], "sparseMatrix")

6 x 18 sparse Matrix of class "dgCMatrix"

A 1 1 1 . . . . . . . . . . . . . . . . . .
B . . . 1 1 1 . . . . . . . . . . . . . . .
C . . . . . . 1 1 1 . . . . . . . . . . . .
D . . . . . . . . . 1 1 1 . . . . . . . . .
E . . . . . . . . . . . . 1 1 1 . . . . . .
F . . . . . . . . . . . . . . . . . 1 1 1
```

The elements represented as '.' in the output are known to be zero and are not stored explicitly.

The `L` component of this fitted model is a Cholesky factorization of a matrix $\boldsymbol{A}^*(\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is a parameter vector determining $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. This matrix can be factored as $\boldsymbol{\Sigma} = \boldsymbol{TSST}^\top$, where $\boldsymbol{T}$ is a unit, lower triangular matrix (that is, all the elements above the diagonal are zero and all the elements on the diagonal are unity) and $\boldsymbol{S}$ is a diagonal matrix with non-negative elements on the diagonal. The matrix $\boldsymbol{A}^*(\boldsymbol{\theta})$ is

$$\begin{aligned}\boldsymbol{A}^*(\boldsymbol{\theta}) &= \begin{bmatrix} \boldsymbol{Z}^{*\top}\boldsymbol{Z}^* + \boldsymbol{I} & \boldsymbol{Z}^{*\top}\boldsymbol{X} & -\boldsymbol{Z}^{*\top}\boldsymbol{y} \\ \boldsymbol{X}^\top\boldsymbol{Z}^* & \boldsymbol{X}^\top\boldsymbol{X} & -\boldsymbol{X}^\top\boldsymbol{y} \\ -\boldsymbol{y}^\top\boldsymbol{Z}^* & -\boldsymbol{y}^\top\boldsymbol{X} & \boldsymbol{y}^\top\boldsymbol{y} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{T}^\top\boldsymbol{S} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & 1 \end{bmatrix} \boldsymbol{A} \begin{bmatrix} \boldsymbol{ST} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & 1 \end{bmatrix} + \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & 0 \end{bmatrix}. \end{aligned} \tag{3}$$

For model (1) the matrices $\boldsymbol{T}$ and $\boldsymbol{S}$ are particularly simple, $\boldsymbol{T} = \boldsymbol{I}_6$, the 6×6 identity matrix and $\boldsymbol{S} = s_{1,1}\boldsymbol{I}_6$ where $s_{1,1} = \sigma_b/\sigma$ is the standard deviation of the random effects relative to the standard deviation of the per-observation noise term $\boldsymbol{\epsilon}$.

The Cholesky decomposition of $\boldsymbol{A}^*$ is a lower triangular sparse matrix $\boldsymbol{L}$

```
> as(Rm1ML@L, "sparseMatrix")
6 x 6 sparse Matrix of class "dtCMatrix"

[1,] 9.797 .     .     .     .     .
[2,] .     9.797 .     .     .     .
[3,] .     .     9.797 .     .     .
[4,] .     .     .     9.797 .     .
[5,] .     .     .     .     9.797 .
[6,] .     .     .     .     .     9.797
```

As explained in later sections the matrix $\boldsymbol{L}$ provides all the information needed to evaluate the ML deviance or the REML deviance as a function of $\boldsymbol{\theta}$. The components of the deviance are given in the `deviance` slot of the fitted model

```
> Rm1ML@deviance
        ML       REML       ldL2      ldRX2    sigmaML  sigmaREML      pwrss
128.560037 122.237086  27.385123 -1.673815   4.020779   4.137348 290.999999
      disc       usqr       wrss        dev       llik    NULLdev
195.010579  95.989420 195.010579         NA         NA         NA
```

The element labelled `ldL2` is the logarithm of the square of the determinant of the upper left 6×6 section of $\boldsymbol{L}$. This corresponds to $\log\left|\boldsymbol{Z}^{*\top}\boldsymbol{Z}^* + \boldsymbol{I}_q\right|$ where $\boldsymbol{Z}^* = \boldsymbol{ZTS}$. We can verify that the value 27.38292 can indeed be calculated in this way.

```
> L <- as(Rm1ML@L, "sparseMatrix")
> 2 * sum(log(diag(L)))
[1] 27.38512
```

The `lr2` element of the `deviance` slot is the logarithm of the penalized residual sum of squares. It can be calculated as the logarithm of the square of the last diagonal element in $\boldsymbol{L}$.

```
> (RX <- Rm1ML@RX)
          [,1]
[1,] 0.4330476
```

For completeness we mention that the `ldRX2` element of the `deviance` slot is the logarithm of the product of the squares of the diagonal elements of $\boldsymbol{L}$ corresponding to columns of $\boldsymbol{X}$.

```
> 2 * sum(log(diag(Rm1ML@RX)))
[1] -1.673815
```

This element is used in the calculation of the REML criterion.

Another slot in the fitted model object is `dims`, which contains information on the dimensions of the model and some of the characteristics of the fit.

```
> (dd <- Rm1ML@dims)
   nt     n     p     q     s    np   LMM  REML  fTyp  lTyp  vTyp  nest useSc
    1    18     1     6     1     1     0     0     2     5     1     1     1
 nAGQ  verb  mxit  mxfn   cvg
    1     1   300   900     5
```

We can reconstruct the ML estimate of the residual variance as the penalized residual sum of squares divided by the number of observations.

```
> Rm1ML@deviance["pwrss"]/dd["n"]
   pwrss
16.16667
```

The *profiled deviance* function

$$\begin{aligned}\tilde{\mathcal{D}}(\boldsymbol{\theta}) &= \log\left|\boldsymbol{Z}^{*\mathsf{T}}\boldsymbol{Z}^{*}+\boldsymbol{I}_q\right| + n\log\left(1+\frac{2\pi r^2}{n}\right)\\ &= n\left[1+\log\left(\frac{2\pi}{n}\right)\right] + \log\left|\boldsymbol{Z}^{*\mathsf{T}}\boldsymbol{Z}^{*}+\boldsymbol{I}_q\right| + n\log r^2\end{aligned} \tag{4}$$

is a function of $\boldsymbol{\theta}$ only. In this case $\boldsymbol{\theta} = \sigma_1$, the relative standard deviation of the random effects, is one-dimensional. The maximum likelihood estimate (mle) of $\boldsymbol{\theta}$ minimizes the profiled deviance. The mle's of all the other parameters in the model can be derived from the estimate of this parameters.

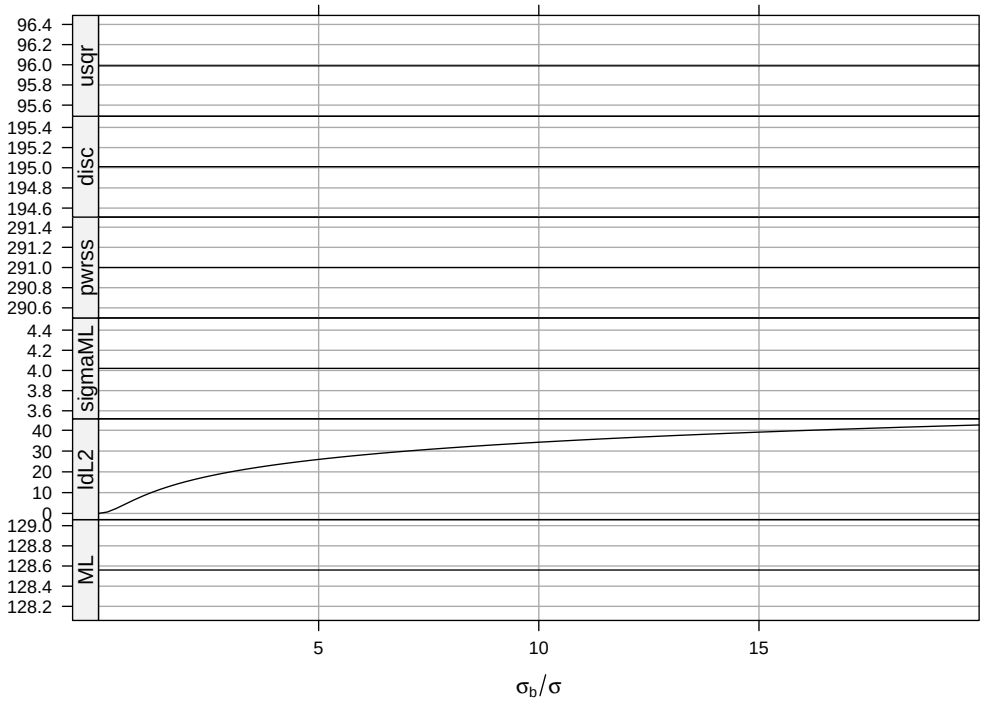

Figure 2: The profiled deviance, and those components of the profiled deviance that vary with $\boldsymbol{\theta}$, as a function of $\boldsymbol{\theta}$ in model `Rm1ML` for the `Rail` data. In this model the parameter $\boldsymbol{\theta}$ is the scalar σ_1, the standard deviation of the random effects relative to the standard deviation of the per-observation noise.

The term $n\left[1+\log\left(2\pi/n\right)\right]$ in (4) does not depend on $\boldsymbol{\theta}$. The other two terms, $\log\left|\boldsymbol{Z}^{*\mathsf{T}}\boldsymbol{Z}^{*}+\boldsymbol{I}_q\right|$ and $n\log r^2$, measure the complexity of the model and the fidelity of the fitted values to the observed data, respectively. We plot the value of each of the varying terms versus σ_1 in Figure 2.

The component $\log\left|\boldsymbol{S}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}\boldsymbol{S}+\boldsymbol{I}\right|$ has the value 0 at $\sigma_1=0$ and increases as σ_1 increases. It is unbounded as $\sigma_1\to\infty$. The component $n\log\left(r^2\right)$ has a finite value at $\sigma_1=0$ from which it decreases as σ_1 increases. The value at $\sigma_1=0$ corresponds to the residual sum of squares for the regression of $\boldsymbol{y}$ on the columns of $\boldsymbol{X}$.

```
> 18 * log(deviance(lm(travel ~ 1, Rail)))
[1] 164.8714
```

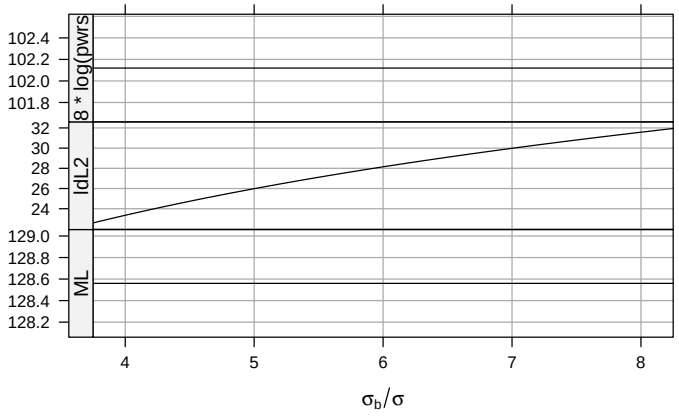

Figure 3: The part of the deviance that varies with σ_1 as a function of σ_1 near the optimum. The component $\log\left|\boldsymbol{S}\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{S}+\boldsymbol{I}\right|$ is shown at the bottom of the frame. This is the component of the deviance that increases with σ_1. Added to this component is $n\log\left[r^2(\sigma_1)\right]-n\log\left[r^2(\infty)\right]$, the comonent of the deviance that decreases as σ_1 increases. Their sum is minimized at $\widehat{\sigma}_1=5.626$.

As $\sigma_1\to\infty$, $n\log\left(r^2\right)$ approaches the value corresponding to the residual sum of squares for the regression of $\boldsymbol{y}$ on the columns of $\boldsymbol{X}$ and $\boldsymbol{Z}$. For this model that is

```
> 18 * log(deviance(lm(travel ~ Rail, Rail)))
[1] 94.82145
```

2 Multiple random effects per level

The `sleepstudy` data are an example of longitudinal data. That is, they are repeated measurements taken on the same experimental units over time. A plot of reaction time versus days of sleep deprivation by subject (Figure 4) shows there is considerable variation between subjects in both the intercept and the slope of the linear trend.

The model

```
> print(sm1 <- lmer(Reaction ~ Days + (Days | Subject), sleepstudy))
```

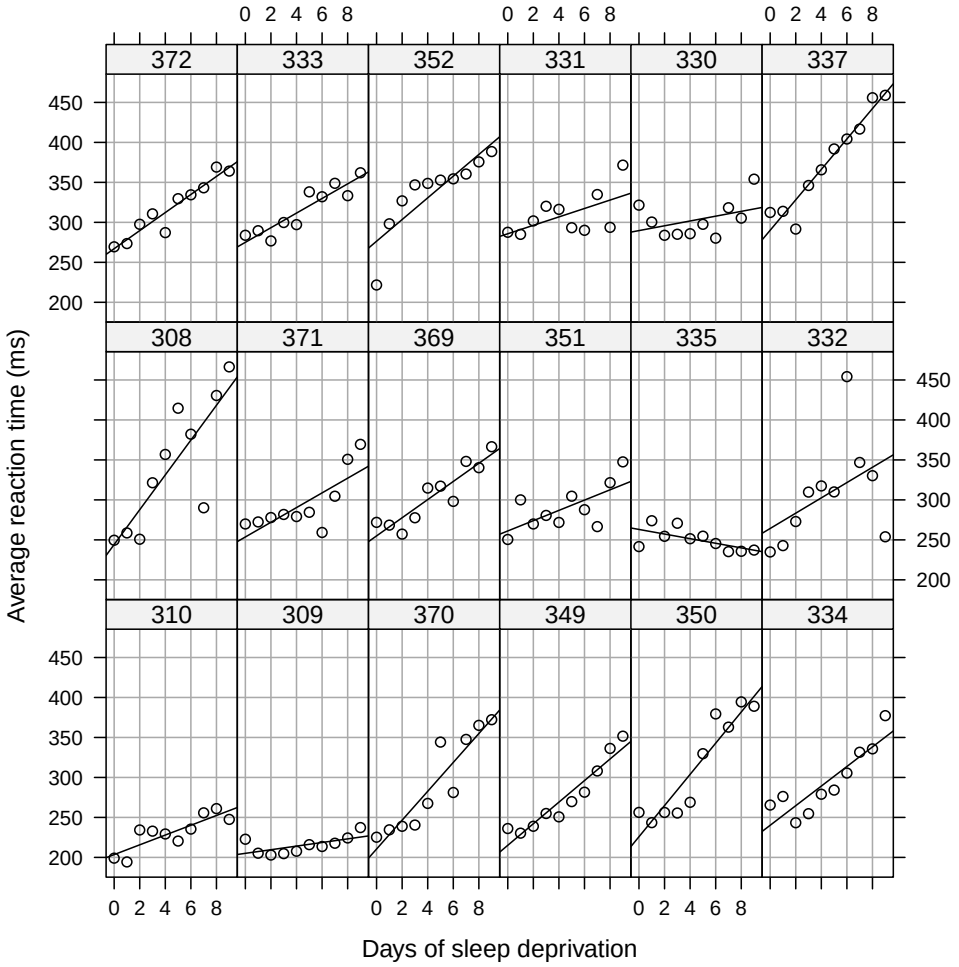

Figure 4: A lattice plot of the average reaction time versus number of days of sleep deprivation by subject for the `sleepstudy` data. Each subject's data are shown in a separate panel, along with a simple linear regression line fit to the data in that panel. The panels are ordered, from left to right along rows starting at the bottom row, by increasing intercept of these per-subject linear regression lines. The subject number is given in the strip above the panel.

```
Linear mixed model fit by REML
Formula: Reaction ~ Days + (Days | Subject)
   Data: sleepstudy
  AIC  BIC logLik deviance REMLdev
 1756 1775 -871.8     1752    1744
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 612.092  24.7405
          Days         35.072   5.9221  0.066
 Residual             654.941  25.5918
Number of obs: 180, groups: Subject, 18

Fixed effects:
            Estimate Std. Error t value
(Intercept)  251.405      6.825   36.84
Days          10.467      1.546    6.77

Correlation of Fixed Effects:
     (Intr)
Days -0.138
```

provides for fixed effects for the intercept (the typical reaction time in the population after zero days of sleep deprivation) and the slope with respect to `Days` (the typical change in reaction time per day of sleep deprivation). In addition there are random effects per subject for both the intercept and the slope parameters.

With two random effects per subject the matrix $\boldsymbol{\Sigma}$ for this model is 36×36 with 18 2×2 diagonal blocks. The matrix $\boldsymbol{A}$ is 39×39 as is $\boldsymbol{L}$, the Cholesky factor of $\boldsymbol{A}^*$. The structure of $\boldsymbol{A}$ and of $\boldsymbol{L}$ are shown in Figure 5. For this model (as for all models with a single random effects expression) the structure of $\boldsymbol{L}$ is identical to that of the lower triangle of $\boldsymbol{A}$.

Like the `Rail` data, the `sleepstudy` data are balanced in that each subject's reaction time is measured the same number of times and at the same times. One consequence of the balance is that the diagonal blocks in the first 36 rows of $\boldsymbol{A}$ are identical as are those in the first 36 rows of $\boldsymbol{L}$.

```
> as(sm1@L, "sparseMatrix")[1:2, 1:2]

2 x 2 sparse Matrix of class "dtCMatrix"

[1,] 3.424559 .
[2,] 3.224091 2.408562

> sm1@RX

        [,1]      [,2]
[1,] 3.78601  2.301304
[2,] 0.00000 16.555992
```

The determinant quantities in

```
> sm1@deviance
```

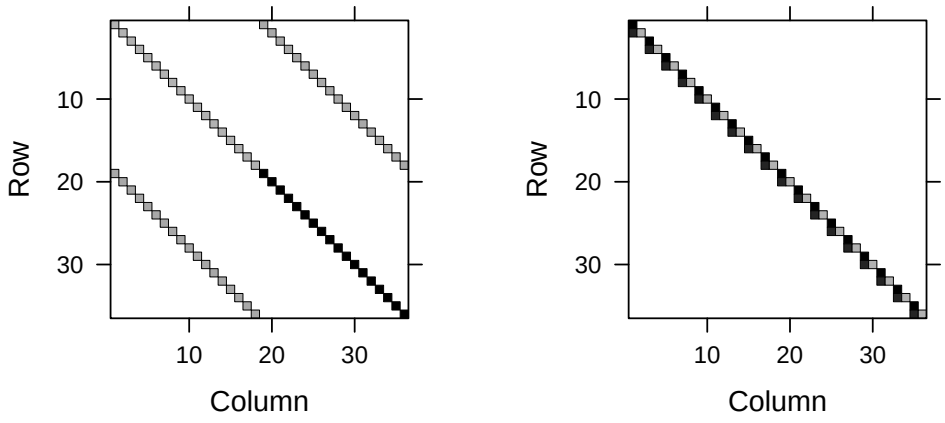

Figure 5: Structure of the sparse matrices $\boldsymbol{A}$ (left panel) and $\boldsymbol{L}$ (right panel) for the model `sm1`. The non-zero elements as depicted as gray squares with larger magnitudes shown as darker gray.

```
          ML         REML         ldL2        ldRX2      sigmaML    sigmaREML
1.751986e+03 1.743628e+03 7.596009e+01 8.276122e+00 2.544924e+01 2.559182e+01
       pwrss         disc         usqr         wrss          dev         llik
1.165795e+05 9.888096e+04 1.769913e+04 9.888039e+04           NA           NA
     NULLdev
          NA
```

are derived from the diagonal elements of $\boldsymbol{L}$. `ldZ` is the logarithm of square of the product of the first 36 elements of the diagonal, `ldX` is the logarithm of the square of the product of the 37th and 38th elements and `lr2` is the logarithm of the square of the 39th element.

```
> sm1@RX
        [,1]      [,2]
[1,] 3.78601  2.301304
[2,] 0.00000 16.555992
> str(dL <- diag(as(sm1@L, "sparseMatrix")))
 num [1:36] 3.42 2.41 3.42 2.41 3.42 ...
> c(ldL2 = sum(log(dL^2)), ldRX2 = sum(log(diag(sm1@RX)^2)), log(sm1@deviance["pwrss"]))
     ldL2     ldRX2     pwrss
75.960094  8.276122 11.666329
```

The 36×36 matrices $\boldsymbol{S}$, $\boldsymbol{T}$ and $\boldsymbol{\Sigma} = \boldsymbol{TSST}^{\mathsf{T}}$ are block-diagonal consisting of 18 identical 2×2 diagonal blocks. The template for the diagonal blocks of $\boldsymbol{S}$ and $\boldsymbol{T}$ is stored as a single matrix

```
> show(st <- sm1@ST[[1]])
```

```
            (Intercept)      Days
(Intercept)  0.96673432 0.0000000
Days         0.01569043 0.2309092
```

where the diagonal elements are those of $\boldsymbol{S}$ and the strict lower triangle is that of $\boldsymbol{T}$.

The `VarCorr` generic function extracts the estimates of the variance-covariance matrices of the random effects. Because model `sm1` has a single random-effects expression there is only one estimated variance-covariance matrix

```
> show(vc <- VarCorr(sm1))
```

```
$Subject
            (Intercept)      Days
(Intercept)  612.091776  9.603985
Days           9.603985 35.071536
attr(,"stddev")
(Intercept)        Days
  24.740489    5.922123
attr(,"correlation")
            (Intercept)       Days
(Intercept)  1.00000000 0.06554896
Days         0.06554896 1.00000000

attr(,"sc")
sigmaREML
 25.59182
```

The `"sc"` attribute of this matrix is the estimate of σ, the standard deviation of the per-observation noise term.

We can reconstruct this variance-covariance estimate as

```
> T <- st
> diag(T) <- 1
> S <- diag(diag(st))
> T
```

```
            (Intercept) Days
(Intercept)  1.00000000    0
Days         0.01569043    1
```

```
> S
```

```
          [,1]      [,2]
[1,] 0.9667343 0.0000000
[2,] 0.0000000 0.2309092
```

```
> T %*% S %*% S %*% t(T) * attr(vc, "sc")^2
```

```
            (Intercept)      Days
(Intercept)  612.091776  9.603985
Days           9.603985 35.071536
```

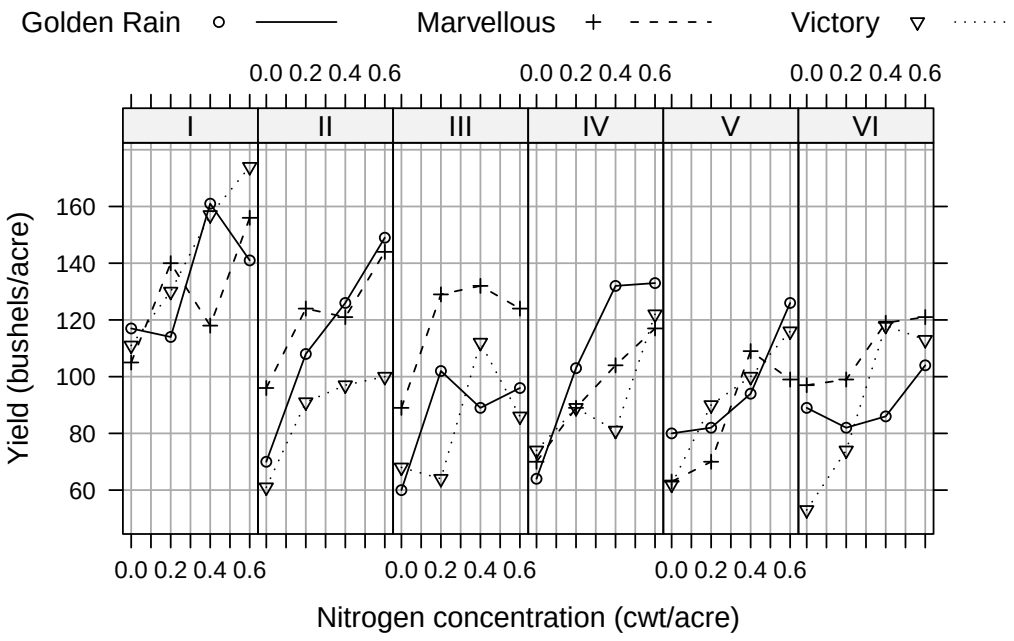

Figure 6: Yield of oats versus applied concentration of nitrogen fertilizer for three different varieties of oats in 6 different locations.

3 A model with two nested grouping factors

The `Oats` data from the `nlme` package came from an experiment in which fields in 6 different locations (the `Block` factor) were each divided into three plots and each of these 18 plots was further subdivided into four subplots. Three varieties of oats were assigned randomly to the three plots in each block and four concentrations of fertilizer (measured as nitrogen concentration) were randomly assigned to the subplots in each plot. The yield on each subplot is the response shown in Figure 6.

The fitted model `Om1`

```
> print(Om1 <- lmer(yield ~ nitro + Variety + (1 | Block/Variety),
+     Oats), corr = FALSE)
Linear mixed model fit by REML
Formula: yield ~ nitro + Variety + (1 | Block/Variety)
   Data: Oats
   AIC   BIC logLik deviance REMLdev
 592.9 608.8 -289.4    601.3   578.9
Random effects:
 Groups        Name        Variance Std.Dev.
```

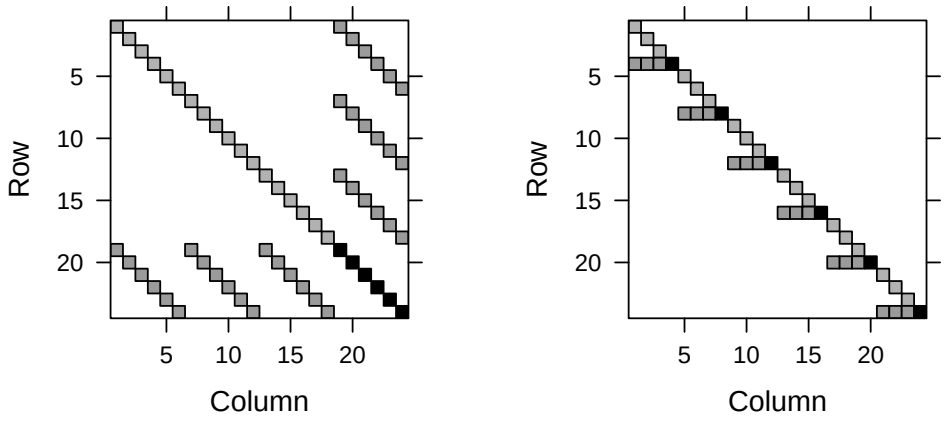

Figure 7: Structure of the sparse matrices $\boldsymbol{A}$ (left panel) and $\boldsymbol{L}$ (right panel) for the model `Om1`.

```
 Variety:Block (Intercept) 108.94   10.438
 Block         (Intercept) 214.48   14.645
 Residual                  165.56   12.867
Number of obs: 72, groups: Variety:Block, 18; Block, 6

Fixed effects:
                  Estimate Std. Error t value
(Intercept)         82.400      8.058  10.226
nitro               73.667      6.781  10.863
VarietyMarvellous    5.292      7.079   0.748
VarietyVictory      -6.875      7.079  -0.971
```

provides fixed effects for the nitrogen concentration and for the varieties (coded as differences relative to the reference variety "Golden Rain") and random effects for each block and for each plot within the block. In this case a plot can be indexed by the combination of variety and block, denoted `Variety:Block` in the output. Notice that there are 18 levels of this grouping factor corresponding to the 18 unique combinations of variety and block.

A given plot occurs in one and only one block. We say that the plot grouping factor is *nested within* the block grouping factor. The structure of the matrices $\boldsymbol{A}$ and $\boldsymbol{L}$ for this model (Figure 7) In the matrix $\boldsymbol{A}$ the first 18 rows and columns correspond to the 18 random effects (1 random effect for each of the 18 levels of this grouping factor). The next 6 rows and columns correspond to the 6 random effects for block (6 levels and 1 random effect

per level). The off-diagonal elements in rows 19 to 24 and columns 1 to 18 indicate which plots and blocks are observed simultaneously. Because the plot grouping factor is nested within the block grouping factor there will be exactly one nonzero in the rows 19 to 24 for each of the columns 1 to 18.

For this model the fixed-effects specification includes indicator vectors with systematic zeros. These appear as systematic zeros in rows 27 and 28 of $\boldsymbol{A}$ and $\boldsymbol{L}$. The statistical analysis of model `Om1` indicates that the systematic effect of the `Variety` factor is not significant and we could omit it from the model, leaving us with

```
> print(Om1a <- lmer(yield ~ nitro + (1 | Block/Variety), Oats),
+     corr = FALSE)
Linear mixed model fit by REML
Formula: yield ~ nitro + (1 | Block/Variety)
   Data: Oats
 AIC   BIC logLik deviance REMLdev
 603 614.4 -296.5    604.3     593
Random effects:
 Groups        Name        Variance Std.Dev.
 Variety:Block (Intercept) 121.10   11.005
 Block         (Intercept) 210.42   14.506
 Residual                  165.56   12.867
Number of obs: 72, groups: Variety:Block, 18; Block, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)   81.872      6.945   11.79
nitro         73.667      6.781   10.86
```

with matrices $\boldsymbol{A}$ and $\boldsymbol{L}$ whose patterns are shown in Figure 8.

In Figures 7 and 8 the pattern in $\boldsymbol{L}$ is different from that of the lower triangle of $\boldsymbol{A}$ but only because a permutation has been applied to the rows and columns of $\boldsymbol{A}^*$ before computing the Cholesky decomposition. The effect of this permutation is to isolate connected blocks of rows and columns close to the diagonal.

The isolation of connected blocks close to the diagonal is perhaps more obvious when the model multiple random-effects expressions based on the same grouping factor. This construction is used to model independent random effects for each level of the grouping factor.

For example, the random effect for the intercept and the random effect for the slope in the sleep-study data could reasonably be modeled as independent, as in the model

```
> print(sm2 <- lmer(Reaction ~ Days + (1 | Subject) + (0 + Days |
+     Subject), sleepstudy), corr = FALSE)
```

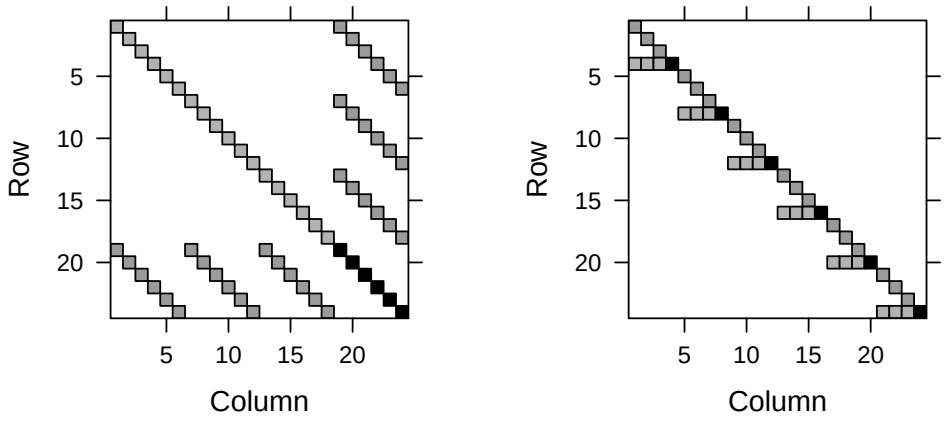

Figure 8: Structure of the sparse matrices $\boldsymbol{A}$ (left panel) and $\boldsymbol{L}$ (right panel) for the model `Om1a`.

```
Linear mixed model fit by REML
Formula: Reaction ~ Days + (1 | Subject) + (0 + Days | Subject)
   Data: sleepstudy
  AIC  BIC logLik deviance REMLdev
 1754 1770 -871.8     1752    1744
Random effects:
 Groups   Name        Variance Std.Dev.
 Subject  (Intercept) 627.568  25.0513
 Subject  Days         35.858   5.9882
 Residual             653.584  25.5653
Number of obs: 180, groups: Subject, 18

Fixed effects:
            Estimate Std. Error t value
(Intercept)  251.405      6.885   36.51
Days          10.467      1.559    6.71
```

The structures of the matrices $\boldsymbol{A}$ and $\boldsymbol{L}$ for this model are shown in Figure 9.

The first 18 elements of $\boldsymbol{b}$ are the random effects for the intercept for each of the 18 subjects followed by the random effects for the slopes for each of the 18 subjects. The (0-based) permutation vector applied to the rows and columns of $\boldsymbol{A}^*$ before taking the decomposition is

```
> str(sm2@L@perm)
 int [1:36] 0 18 1 19 2 20 3 21 4 22 ...
```

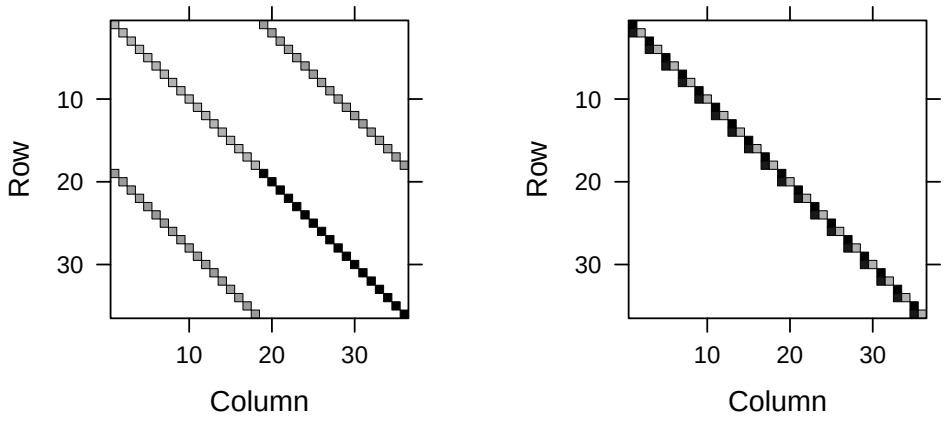

Figure 9: Structure of the sparse matrices $\boldsymbol{A}$ (left panel) and $\boldsymbol{L}$ (right panel) for the model `sm2`.

This means that, in the 1-based indexing system used in R, the permutation will pair up the first and 19th rows and columns, the 2nd and 20th rows and columns, etc. resulting in the pattern for $\boldsymbol{L}$ shown in Figure 9

Figure 6 indicates that the slope of yield versus nitrogen concentration may depend on the block but not the plot within the block. We can fit such a model as

```
> print(Om2 <- lmer(yield ~ nitro + (1 | Variety:Block) + (nitro |
+     Block), Oats), corr = FALSE)
Linear mixed model fit by REML
Formula: yield ~ nitro + (1 | Variety:Block) + (nitro | Block)
   Data: Oats
   AIC   BIC logLik deviance REMLdev
 606.8 622.7 -296.4    604.1   592.8
Random effects:
 Groups        Name        Variance Std.Dev. Corr
 Variety:Block (Intercept) 121.066  11.0030
 Block         (Intercept) 177.434  13.3204
               nitro        15.867   3.9833  1.000
 Residual                  164.661  12.8320
Number of obs: 72, groups: Variety:Block, 18; Block, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)   81.872      6.535   12.53
nitro         73.667      6.956   10.59
```

The structures of the matrices $\boldsymbol{A}$ and $\boldsymbol{L}$ for this model are shown in Figure 10.

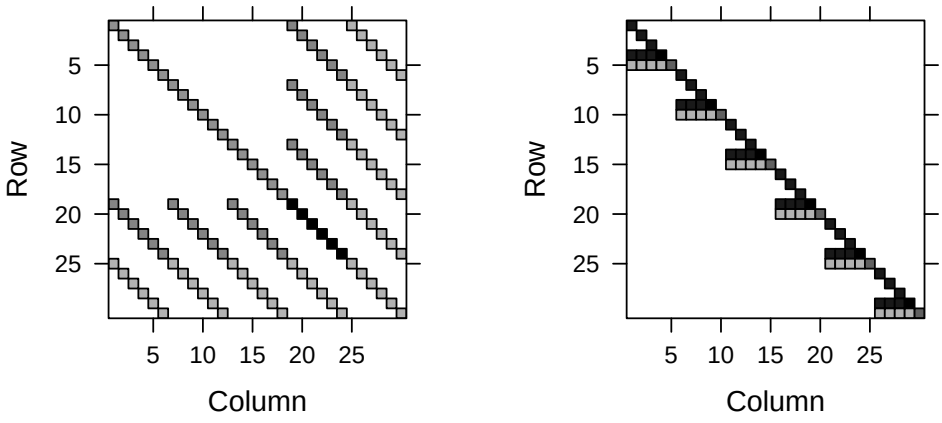

Figure 10: Structure of the sparse matrices $\boldsymbol{A}$ (left panel) and $\boldsymbol{L}$ (right panel) for the model `Om2`.

We see that the only difference in the structure of the $\boldsymbol{A}$ matrices from models `Om1a` and `Om2` is that rows and columns 19 to 24 from model `Om1a` have been replicated. Thus the 1×1 blocks on the diagonal of $\boldsymbol{A}$ in positions 19 to 24 for model `Om1a` become 2×2 blocks in model `Om2`.

This replication of rows associated with levels of the `Block` factor carries over to the matrix $\boldsymbol{L}$.

The property of being nested or not is often attributed to random effects. In fact, nesting is a property of the grouping factors with whose levels the random effects are associated. In both models `Om1a` and `Om2` the `Variety:Block` factor is nested within `Block`. If the grouping factors in the random effects terms in a model form a nested sequence then the matrix $\boldsymbol{A}^*$ and its Cholesky decomposition $\boldsymbol{L}$ will have the property that the number of nonzeros in $\boldsymbol{L}$ is the same as the number of nonzeros in the lower triangle of $\boldsymbol{A}^*$. That is, there will be no "fill-in" generating new nonzero positions when forming the Cholesky decomposition.

To check this we can examine the number of nonzero elements in $\boldsymbol{A}$ and $\boldsymbol{L}$ for these models. Because the matrix $\boldsymbol{A}$ is stored as a symmetric matrix with only the non-redundant elements being stored explicitly, the number of stored nonzeros in these two matrices are identical.

```
> length(tcrossprod(Om2@A)@x)
[1] 72
```

```
> length(Om2@L@x)
[1] 72
```

4 Non-nested grouping factors

When grouping factors are not nested they are said to be "crossed". Sometimes we will distinguish between `partially crossed` grouping factors and `completely crossed` grouping factors. When two grouping factors are completely crossed, every level of the first factor occurs at least once with every level of the second factor - creating matrices $\boldsymbol{A}$ and $\boldsymbol{L}$ with dense off-diagonal blocks.

In observational studies it is more common to encounter partially crossed grouping factors. For example, the `ScotsSec` data in the `mlmRev` package provides the attainment scores of 3435 students in Scottish secondary schools as well as some demographic information on the students and an indicator of which secondary school and which primary school the student attended.

```
> str(ScotsSec)
'data.frame':        3435 obs. of  6 variables:
 $ verbal : num  11 0 -14 -6 -30 -17 -17 -11 -9 -19 ...
 $ attain : num  10 3 2 3 2 2 4 6 4 2 ...
 $ primary: Factor w/ 148 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ sex    : Factor w/ 2 levels "M","F": 1 2 1 1 2 2 2 1 1 1 ...
 $ social : num  0 0 0 20 0 0 0 0 0 0 ...
 $ second : Factor w/ 19 levels "1","2","3","4",..: 9 9 9 9 9 9 1 1 9 9 ...
```

If we use both `primary` and `second` as grouping factors for random effects in a model the only possibility for these factors to form a nested sequence is to have `primary` nested within `second` (because there are 148 levels of `primary` and 19 levels of `second`). We could check if these are nested by doing a cross-tabulation of these factors but it is easier to fit an initial model

```
> print(Sm1 <- lmer(attain ~ verbal * sex + (1 | primary) + (1 |
+     second), ScotsSec), corr = FALSE)
Linear mixed model fit by REML
Formula: attain ~ verbal * sex + (1 | primary) + (1 | second)
   Data: ScotsSec
   AIC   BIC logLik deviance REMLdev
 14882 14925  -7434    14843   14868
Random effects:
 Groups   Name        Variance Std.Dev.
 primary  (Intercept) 0.275453 0.52484
 second   (Intercept) 0.014747 0.12144
 Residual             4.253114 2.06231
Number of obs: 3435, groups: primary, 148; second, 19

Fixed effects:
```

```
            Estimate Std. Error t value
(Intercept) 5.914728   0.076783   77.03
verbal      0.158356   0.003787   41.81
sexF        0.121552   0.072413    1.68
verbal:sexF 0.002593   0.005388    0.48
```

and examine the "nest" element of the dims slot.

```
> Sm1@dims
```

```
   nt     n     p     q     s    np   LMM  REML  fTyp  lTyp  vTyp  nest useSc
    2  3435     4   167     1     2     0     1     2     5     1     0     1
 nAGQ  verb  mxit  mxfn   cvg
    1     0   300   900     4
```

We see that these grouping factors are not nested. That is, some of the elementary schools sent students to more than one secondary school.

Now that we know the answer we can confirm it by checking the first few rows of the cross-tabulation

```
> head(xtabs(~primary + second, ScotsSec))
```

```
       second
primary  1 2 3 4  5 6 7 8  9 10 11 12 13 14 15 16 17 18 19
      1  8 0 0 0  0 0 0 0 45  0  0  0  0  0  0  0  0  1  0
      2  0 0 0 0  0 0 7 0  0  0  0  0  0  0  0  0  0  0  0
      3  0 0 0 0  3 0 0 0  0  0  0  0  0  0  0  0  0  0  0
      4  0 0 0 0  0 1 0 0  6  0  0  0  0  0  0  0  0  0  0
      5 53 0 0 0  0 0 0 0  0  0  0  0  0  0  0  0  0  0  0
      6  1 0 1 0 52 0 0 0  0  0  1  0  0  0  0  0  0  0  0
```

We see that primary schools 1, 4 and 6 each occurred with multiple secondary schools.

For non-nested grouping factors like these, the structure of $\boldsymbol{A}$ and $\boldsymbol{L}$, shown in Figure 11 is more complex than for nested grouping factors. The matrix $\boldsymbol{A}$ has a 148×148 diagonal block in the upper left, corresponding the the 148 levels of the primary factor, followed on the diagonal by a 19×19 diagonal block corresponding to the 19 levels of the second factor. However, the off-diagonal block in rows 149 to 167 and columns 1 to 148 does not have a simple structure. There is an indication of three groups of primary and secondary schools but even those groups are not exclusive.

With non-nested grouping factors such as these there can be fill-in. That is, the number of nonzeros in $\boldsymbol{L}$ is greater than the number of non-redundant nonzeros in $\boldsymbol{A}$.

```
> c(A = length(tcrossprod(Sm1@A)@x), L = length(Sm1@L@x))
```

```
  A   L
470 594
```

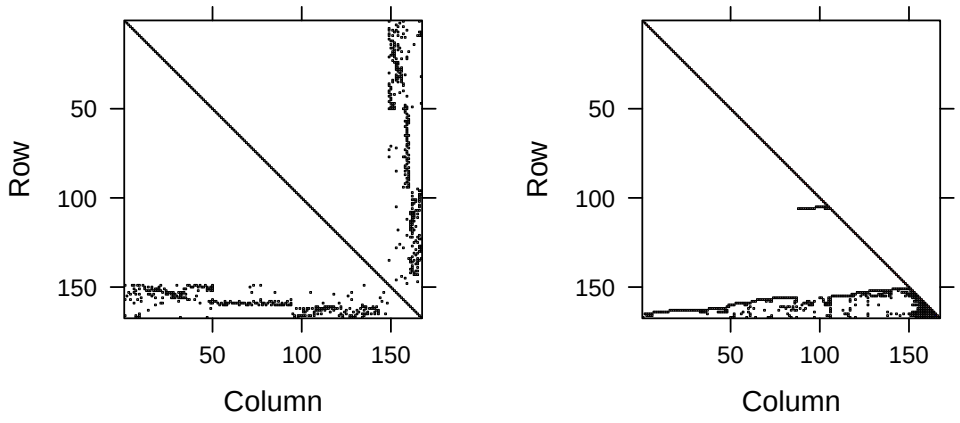

Figure 11: Structure of the sparse matrices $\boldsymbol{A}$ (left panel) and $\boldsymbol{L}$ (right panel) for the model `Sm1`.

The permutation applied to the rows and columns of $\boldsymbol{A}$ is a "fill-reducing" permutation chosen to reduce the amount of fill-in during the Cholesky decomposition. The approximate minimal degree (AMD) algorithm (Davis, 2006) is used to select this permutation when non-nested grouping factors are detected. It is followed by a "post-ordering" permutation that isolates connected blocks on the diagonal.

5 Structure of Σ and Z

The columns of $\boldsymbol{Z}$ and the rows and columns of $\boldsymbol{\Sigma}$ are associated with the levels of one or more grouping factors in the data. For example, a common application of linear mixed models is the analysis of students' scores on the annual state-wide performance tests mandated by the No Child Left Behind Act. A given score is associated with a student, a teacher, a school and a school district. These could all be grouping factors in a model.

We write the grouping factors as $\boldsymbol{f}_i, i = 1, \ldots k$. The number of levels of the ith factor, $\boldsymbol{f}_i$, is n_i and the number of random effects associated with each level is q_i. For example, if $\boldsymbol{f}_1$ is "student" then n_1 is the number of students in the study. If we have a simple additive random effect for each student

then $q_1 = 1$. If we have a random effect for both the intercept and the slope with respect to time for each student then $q_1 = 2$. The $q_i, i = 1, \ldots, k$ are typically very small whereas the $n_i, i = 1, \ldots, k$ can be very large.

In the statistical model we assume that random effects associated with different grouping factors are independent, which implies that $\boldsymbol{\Sigma}$ is block diagonal with k diagonal blocks of sizes $n_i q_i \times n_i q_i, i = 1, \ldots, k$. That is

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2 & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{\Sigma}_k \end{bmatrix} \tag{5}$$

Furthermore, random effects associated with different levels of the same grouping factor are assumed to be independent and identically distributed, which implies that $\boldsymbol{\Sigma}_i$ is itself block diagonal in n_i blocks and that each of these blocks is a copy of a $q_i \times q_i$ matrix $\tilde{\boldsymbol{\Sigma}}_i$. That is

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \tilde{\boldsymbol{\Sigma}}_i & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \tilde{\boldsymbol{\Sigma}}_i & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \tilde{\boldsymbol{\Sigma}}_i \end{bmatrix} = \boldsymbol{I}_{n_i} \otimes \tilde{\boldsymbol{\Sigma}}_i \quad i = 1, \ldots, k \tag{6}$$

where $\otimes$ denotes the Kronecker product.

The condition that $\boldsymbol{\Sigma}$ is positive semi-definite holds if and only if the $\tilde{\boldsymbol{\Sigma}}_i, i = 1, \ldots, k$ are positive semi-definite. To ensure that the $\tilde{\boldsymbol{\Sigma}}_i$ are positive semi-definite, we express them as

$$\tilde{\boldsymbol{\Sigma}}_i = \tilde{\boldsymbol{T}}_i \tilde{\boldsymbol{S}}_i \tilde{\boldsymbol{S}}_i \tilde{\boldsymbol{T}}_i^{\mathsf{T}}, \quad i = 1, \ldots, k \tag{7}$$

where $\tilde{\boldsymbol{T}}_i$ is a $q_i \times q_i$ unit lower-triangular matrix (i.e. all the elements above the diagonal are zero and all the diagonal elements are unity) and $\tilde{\boldsymbol{S}}_i$ is a $q_i \times q_i$ diagonal matrix with non-negative elements on the diagonal.

This is the "LDL" form of the Cholesky decomposition of positive semi-definite matrices except that we express the diagonal matrix $\boldsymbol{D}$, which is on the variance scale, as the square of the diagonal matrix $\boldsymbol{S}$, which is on the standard deviation scale. The profiled deviance behaves more like a quadratic on the standard deviation scale than it does on the variance scale so the use of the standard deviation scale enhances convergence.

The $n_i q_i \times n_i q_i$ matrices $\boldsymbol{S}_i, \boldsymbol{T}_i$, $i = 1, \ldots, k$ and the $q \times q$ matrices $\boldsymbol{S}$ and $\boldsymbol{T}$ are defined analogously to (6) and (5). In particular,

$$\boldsymbol{S}_i = \boldsymbol{I}_{n_i} \otimes \tilde{\boldsymbol{S}}_i, \quad i = 1, \ldots, k \tag{8}$$

$$\boldsymbol{T}_i = \boldsymbol{I}_{n_i} \otimes \tilde{\boldsymbol{T}}_i, \quad i = 1, \ldots, k \tag{9}$$

Note that when $q_i = 1$, $\tilde{\boldsymbol{T}}_i = \boldsymbol{I}$ and hence $\boldsymbol{T}_i = \boldsymbol{I}$. Furthermore, $\boldsymbol{S}_i$ is a multiple of the identity matrix in this case.

The parameter vector $\boldsymbol{\theta}_i, i = 1, \ldots, k$ consists of the q_i diagonal elements of $\tilde{\boldsymbol{S}}_i$, which are constrained to be non-negative, followed by the $q_i(q_i - 1)/2$ elements in the strict lower triangle of $\tilde{\boldsymbol{T}}_i$ (in column-major ordering). These last $q_i(q_i - 1)/2$ elements are unconstrained. The $\boldsymbol{\theta}_i$ are combined as

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \\ \vdots \\ \boldsymbol{\theta}_k \end{bmatrix}.$$

Each of the $q \times q$ matrices $\boldsymbol{S}$, $\boldsymbol{T}$ and $\boldsymbol{\Sigma}$ in the decomposition $\boldsymbol{\Sigma} = \boldsymbol{T}\boldsymbol{S}\boldsymbol{S}\boldsymbol{T}^{\mathsf{T}}$ is a function of $\boldsymbol{\theta}$.

As a unit triangular matrix $\boldsymbol{T}$ is non-singular. That is, $\boldsymbol{T}^{-1}$ exists and is easily calculated from the $\tilde{\boldsymbol{T}}_i^{-1}, i = 1, \ldots, k$. When $\boldsymbol{\theta}$ is not on the boundary defined by the constraints, $\boldsymbol{S}$ is a diagonal matrix with strictly positive elements on the diagonal, which implies that $\boldsymbol{S}^{-1}$ exists and that $\boldsymbol{\Sigma}$ is non-singular with $\boldsymbol{\Sigma}^{-1} = \boldsymbol{T}^{-\mathsf{T}}\boldsymbol{S}^{-1}\boldsymbol{S}^{-1}\boldsymbol{T}^{-1}$.

When $\boldsymbol{\theta}$ is on the boundary the matrices $\boldsymbol{S}$ and $\boldsymbol{\Sigma}$ exist but are not invertible. We say that $\boldsymbol{\Sigma}$ is a *degenerate* variance-covariance matrix in the sense that one or more linear combinations of the vector $\boldsymbol{b}$ are defined to have zero variance. That is, the distribution of these linear combinations is a point mass at 0.

The maximum likelihood estimates of $\boldsymbol{\theta}$ (or the restricted maximum likelihood estimates, defined below) can be located on the boundary. That is, they can correspond to a degenerate variance-covariance matrix and we must be careful to allow for this case. However, to begin we consider the non-degenerate case.

6 Methods for non-singular $\boldsymbol{\Sigma}$

When $\boldsymbol{\theta}$ is not on the boundary we can define a standardized random effects vector

$$\boldsymbol{b}^* = \boldsymbol{S}^{-1}\boldsymbol{T}^{-1}\boldsymbol{b} \tag{10}$$

with the properties

$$\begin{aligned}
\mathsf{E}[\boldsymbol{b}^*] &= \boldsymbol{S}^{-1}\boldsymbol{T}^{-1}\mathsf{E}[\boldsymbol{b}] && (11)\\
\mathsf{Var}[\boldsymbol{b}^*] = \boldsymbol{0} &= \mathsf{E}[\boldsymbol{b}^*\boldsymbol{b}^{*\mathsf{T}}] && (12)\\
&= \boldsymbol{S}^{-1}\boldsymbol{T}^{-1}\mathsf{Var}[\boldsymbol{b}]\boldsymbol{T}^{-\mathsf{T}}\boldsymbol{S}^{-1}\\
&= \sigma^2\boldsymbol{S}^{-1}\boldsymbol{T}^{-1}\boldsymbol{\Sigma}\boldsymbol{T}^{-\mathsf{T}}\boldsymbol{S}^{-1}\\
&= \sigma^2\boldsymbol{S}^{-1}\boldsymbol{T}^{-1}\boldsymbol{T}\boldsymbol{S}\boldsymbol{S}\boldsymbol{T}^{\mathsf{T}}\boldsymbol{T}^{-\mathsf{T}}\boldsymbol{S}^{-1}\\
&= \sigma^2\boldsymbol{I}.
\end{aligned}$$

Thus, the unconditional distribution of the q elements of $\boldsymbol{b}^*$ is $\boldsymbol{b}^* \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, like that of the n elements of $\boldsymbol{\epsilon}$.

Obviously the transformation from $\boldsymbol{b}^*$ to $\boldsymbol{b}$ is

$$\boldsymbol{b} = \boldsymbol{T}\boldsymbol{S}\boldsymbol{b}^* \tag{13}$$

and the $n \times q$ model matrix for $\boldsymbol{b}^*$ is

$$\boldsymbol{Z}^* = \boldsymbol{Z}\boldsymbol{T}\boldsymbol{S} \tag{14}$$

so that

$$\boldsymbol{Z}^*\boldsymbol{b}^* = \boldsymbol{Z}\boldsymbol{T}\boldsymbol{S}\boldsymbol{S}^{-1}\boldsymbol{T}^{-1}\boldsymbol{b} = \boldsymbol{Z}\boldsymbol{b}. \tag{15}$$

Notice that $\boldsymbol{Z}^*$ can be evaluated even when $\boldsymbol{\theta}$ is on the boundary. Also, if we have a value of $\boldsymbol{b}^*$ in such a case, we can evaluated $\boldsymbol{b}$ from $\boldsymbol{b}^*$.

Given the data $\boldsymbol{y}$ and values of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, the mode of the conditional distribution of $\boldsymbol{b}^*$ is the solution to a penalized least squares problem

$$\begin{aligned}
\tilde{\boldsymbol{b}}^*(\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{y}) &= \arg\min_{\boldsymbol{b}^*} \left[\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}^*\boldsymbol{b}^*\|^2 + \boldsymbol{b}^{*\mathsf{T}}\boldsymbol{b}^* \right]\\
&= \arg\min_{\boldsymbol{b}^*} \left\| \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix} - \begin{bmatrix} \boldsymbol{Z}^* & \boldsymbol{X} \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{b}^* \\ \boldsymbol{\beta} \end{bmatrix} \right\|^2 .
\end{aligned} \tag{16}$$

In fact, if we optimize the penalized least squares expression in (16) with respect to both $\boldsymbol{b}$ and $\boldsymbol{\beta}$ we obtain the conditional estimates $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}|\boldsymbol{y})$ and the

conditional modes $\tilde{\boldsymbol{b}}^*(\boldsymbol{\theta}, \widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})|\boldsymbol{y})$) which we write as $\widehat{\boldsymbol{b}}^*(\boldsymbol{\theta})$. That is,

$$\begin{aligned}\begin{bmatrix}\widehat{\boldsymbol{b}}^*(\boldsymbol{\theta})\\ \widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\end{bmatrix} &= \arg\min_{\boldsymbol{b}^*,\boldsymbol{\beta}} \left\| \begin{bmatrix}\boldsymbol{Z}^* & \boldsymbol{X} & -\boldsymbol{y}\\ \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0}\end{bmatrix} \begin{bmatrix}\boldsymbol{b}^*\\ \boldsymbol{\beta}\\ 1\end{bmatrix} \right\|^2 \\ &= \arg\min_{\boldsymbol{b}^*,\boldsymbol{\beta}} \begin{bmatrix}\boldsymbol{b}^*\\ \boldsymbol{\beta}\\ 1\end{bmatrix}^\mathsf{T} \boldsymbol{A}^*(\boldsymbol{\theta}) \begin{bmatrix}\boldsymbol{b}^*\\ \boldsymbol{\beta}\\ 1\end{bmatrix}\end{aligned} \tag{17}$$

where the matrix $\boldsymbol{A}^*(\boldsymbol{\theta})$ is as shown in (3) and

$$\boldsymbol{A} = \begin{bmatrix}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} & \boldsymbol{Z}^\mathsf{T}\boldsymbol{X} & -\boldsymbol{Z}^\mathsf{T}\boldsymbol{y}\\ \boldsymbol{X}^\mathsf{T}\boldsymbol{Z} & \boldsymbol{X}^\mathsf{T}\boldsymbol{X} & -\boldsymbol{X}^\mathsf{T}\boldsymbol{y}\\ -\boldsymbol{y}^\mathsf{T}\boldsymbol{Z} & -\boldsymbol{y}^\mathsf{T}\boldsymbol{X} & \boldsymbol{y}^\mathsf{T}\boldsymbol{y}\end{bmatrix}. \tag{18}$$

Note that $\boldsymbol{A}$ does not depend upon $\boldsymbol{\theta}$. Furthermore, the nature of the model matrices $\boldsymbol{Z}$ and $\boldsymbol{X}$ ensures that the pattern of nonzeros in $\boldsymbol{A}^*(\boldsymbol{\theta})$ is the same as that in $\boldsymbol{A}$.

Let the $q \times q$ permutation matrix $\boldsymbol{P}_Z$ represent a fill-reducing permutation for $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ and $\boldsymbol{P}_X$, of size $p \times p$, represent a fill-reducing permutation for $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$. These could be determined, for example, using the *approximate minimal degree* (AMD) algorithm described in Davis (2006) and Davis (1996) and implemented in both the `Csparse` (Davis, 2005b) and the `CHOLMOD` (Davis, 2005a) libraries of C functions. (In many cases $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$ is dense, but of small dimension compared to $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$, and $\boldsymbol{Z}^\mathsf{T}\boldsymbol{X}$ is nearly dense so $\boldsymbol{P}_X$ can be $\boldsymbol{I}_p$, the $p \times p$ identity matrix.)

Let the permutation matrix $\boldsymbol{P}$ be

$$\boldsymbol{P} = \begin{bmatrix}\boldsymbol{P}_Z & \boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{P}_X & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0} & 1\end{bmatrix} \tag{19}$$

and $\boldsymbol{L}(\boldsymbol{\theta})$ be the sparse Cholesky decomposition of $\boldsymbol{A}^*(\boldsymbol{\theta})$ relative to this permutation. That is, $\boldsymbol{L}(\boldsymbol{\theta})$ is a sparse lower triangular matrix with the property that

$$\boldsymbol{L}(\boldsymbol{\theta})\boldsymbol{L}(\boldsymbol{\theta})^\mathsf{T} = \boldsymbol{P}\boldsymbol{A}^*(\boldsymbol{\theta})\boldsymbol{P}^\mathsf{T} \tag{20}$$

For $\boldsymbol{L}(\boldsymbol{\theta})$ to exist we must ensure that $\boldsymbol{A}^*(\boldsymbol{\theta})$ is positive definite. Examination of (17) shows that this will be true if $\boldsymbol{X}$ is of full column rank and $\boldsymbol{y}$

does not lie in the column span of $\boldsymbol{X}$ (or, in statistical terms, if we can't fit $\boldsymbol{y}$ perfectly using only the fixed effects).

Let $r > 0$ be the last element on the diagonal of $\boldsymbol{L}$. Then the minumum penalized residual sum of squares in (17) is r^2 and it occurs at $\widehat{\boldsymbol{b}^*}(\boldsymbol{\theta})$ and $\hat{\beta}(\boldsymbol{\theta})$, the solutions to the sparse triangular system

$$\boldsymbol{L}(\boldsymbol{\theta})^{\mathsf{T}}\boldsymbol{P}\begin{bmatrix}\widehat{\boldsymbol{b}^*}(\boldsymbol{\theta})\\ \widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\\ 1\end{bmatrix} = \begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{0}\\ r\end{bmatrix} \tag{21}$$

(Technically we should not write the 1 in the solution; it should be an unknown. However, for $\boldsymbol{L}$ lower triangular with r as the last element on the diagonal and $\boldsymbol{P}$ a permutation that does not move the last row, the solution for this "unknown" will always be 1.) Furthermore, $\log|\boldsymbol{Z}^{*\mathsf{T}}\boldsymbol{Z}+\boldsymbol{I}|$ can be evaluated as the sum of the logarithms of the first q diagonal elements of $\boldsymbol{L}(\boldsymbol{\theta})$.

The *profiled deviance function*, $\tilde{\mathcal{D}}(\boldsymbol{\theta})$, which is negative twice the log-likelihood of model (2) evaluated at $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ and $\hat{\sigma}^2(\boldsymbol{\theta})$, can be expressed as

$$\tilde{\mathcal{D}}(\boldsymbol{\theta}) = \log\left|\boldsymbol{Z}^{*\mathsf{T}}\boldsymbol{Z}^{*}+\boldsymbol{I}\right| + n\left(1+\log\frac{2\pi r^2}{n}\right). \tag{22}$$

Notice that it is not necessary to solve for $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ or $\widehat{\boldsymbol{b}^*}(\boldsymbol{\theta})$ or $\widehat{\boldsymbol{b}}(\boldsymbol{\theta})$ to be able to evaluate $d(\boldsymbol{\theta})$. All that is needed is to update $\boldsymbol{A}$ to form $\boldsymbol{A}^*$ from which the sparse Cholesky decomposition $\boldsymbol{L}(\boldsymbol{\theta})$ can be calculated and $\tilde{\mathcal{D}}(\boldsymbol{\theta})$ evaluated.

Once $\hat{\boldsymbol{\theta}}$ is determined we can solve for $\widehat{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}})$ and $\widehat{\boldsymbol{b}^*}(\boldsymbol{\theta})$ using (21) and for

$$\widehat{\sigma}^2(\hat{\boldsymbol{\theta}}) = \frac{r^2(\hat{\boldsymbol{\theta}})}{n}. \tag{23}$$

Furthermore, $\widehat{\boldsymbol{b}}(\hat{\boldsymbol{\theta}}) = \boldsymbol{S}\boldsymbol{T}\widehat{\boldsymbol{b}^*}(\hat{\boldsymbol{\theta}})$.

7 Methods for singular $\boldsymbol{\Sigma}$

When $\boldsymbol{\theta}$ is on the boundary, corresponding to a singular $\boldsymbol{\Sigma}$, some of the columns of $\boldsymbol{Z}^*$ are zero. However, the matrix $\boldsymbol{A}^*$ is non-singular and elements of $\boldsymbol{b}^*$ corresponding to the zeroed columns in $\boldsymbol{Z}^*$ approach zero smoothly as

$\boldsymbol{\theta}$ approaches the boundary. Thus $r(\boldsymbol{\theta})$ and $\left|\boldsymbol{Z}^{*\mathsf{T}}\boldsymbol{Z}+\boldsymbol{I}\right|$ are well-defined, as are $\tilde{\mathcal{D}}(\boldsymbol{\theta})$ and the conditional modes $\widehat{\boldsymbol{b}}(\boldsymbol{\theta})$.

In other words, (3) and (20) can be used to define $\tilde{\mathcal{D}}(\boldsymbol{\theta})$ whether or not $\boldsymbol{\theta}$ is on the boundary.

8 REML estimates

It is common to estimate the per-observation noise variance σ^2 in a fixed-effects linear model as $\hat{\sigma}^2 = r^2/(n-p)$ where r^2 is the (unpenalized) residual sum-of-squares, n is the number of observations and p is the number of fixed-effects parameters. This is not the maximum likelihood estimate of σ^2, which is r^2/n. It is the "restricted" or "residual" maximum likelihood (REML) estimate, which takes into account that the residual vector $\boldsymbol{y}-\hat{\boldsymbol{y}}$ is constrained to a linear subspace of dimension $n-p$ in the response space. Thus its squared length, $\|\boldsymbol{y}-\hat{\boldsymbol{y}}\|^2 = r^2$, has only $n-p$ *degrees of freedom* associated with it.

The profiled REML deviance for a linear mixed model can be expressed as

$$\tilde{\mathcal{D}}_R(\boldsymbol{\theta}) = \log\left|\boldsymbol{Z}^{*\mathsf{T}}\boldsymbol{Z}^{*}+\boldsymbol{I}\right| + \log\left|\boldsymbol{L}_{\boldsymbol{X}}\right|^2 + (n-p)\left(1+\log\frac{2\pi r^2}{n-p}\right). \quad (24)$$

9 Generalized linear mixed models

9.1 Generalized linear models

As described in McCullagh and Nelder (1989), a generalized linear model is a statistical model in which the *linear predictor* for the ith response, $\eta_i = \boldsymbol{x}_i\boldsymbol{\beta}$ where $\boldsymbol{x}_i$ is the ith row of the $n\times p$ model matrix $\boldsymbol{X}$ derived from the form of the model and the values of any covariates, is related to the *expected value of the response*, μ_i, through an invertible *link function*, g. That is

$$\boldsymbol{x}_i\boldsymbol{\beta} = \eta_i = g(\mu_i) \quad i = 1,\ldots,n \quad (25)$$

and

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(\boldsymbol{x}_i\boldsymbol{\beta}) \quad i = 1,\ldots,n \quad (26)$$

When the distribution of y_i given μ_i is from the exponential family there exist a *natural* link function for the family (McCullagh and Nelder, 1989).

For a binomial response the natural link is the *logit* link defined as

$$\eta_i = g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) \quad i = 1, \ldots, n \tag{27}$$

with inverse link

$$\mu_i = g^{-1}(\eta_i) = \frac{1}{1+\exp(-\eta_i)} \quad i = 1, \ldots, n \tag{28}$$

Because μ_i is the probability of the ith observation being a "success", η_i is the log of the odds ratio.

The parameters $\boldsymbol{\beta}$ in a generalized linear model are generally estimated by *iteratively reweighted least squares* (IRLS). At each iteration in this algorithm the current parameter estimates are replaced by the parameter estimates of a weighted least squares fit with model matrix $\boldsymbol{X}$ to an adjusted dependent variable. The weights and the adjusted dependent variable are calculated from the link function and the current parameter values.

9.2 Generalized linear mixed models

In a generalized linear mixed model (GLMM) the n-dimensional vector of linear predictors, $\boldsymbol{\eta}$, incorporates both fixed effects, $\boldsymbol{\beta}$, and random effects, $\boldsymbol{b}$, as

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} \tag{29}$$

where $\boldsymbol{X}$ is an $n \times p$ model matrix and $\boldsymbol{Z}$ is an $n \times q$ model matrix.

As for linear mixed models, we model the distribution of the random effects as a multivariate normal (Gaussian) distribution with mean $\mathbf{0}$ and $q \times q$ variance-covariance matrix $\boldsymbol{\Sigma}$. That is,

$$\boldsymbol{b} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})\right). \tag{30}$$

The maximum likelihood estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ maximize the likelihood of the parameters, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, given the observed data, $\boldsymbol{y}$. This likelihood is numerically equivalent to the marginal density of $\boldsymbol{y}$ given $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, which is

$$f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\theta}) = \int_{\boldsymbol{b}} p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{b}) f(\boldsymbol{b}|\boldsymbol{\Sigma}(\boldsymbol{\theta}))\, d\boldsymbol{b} \tag{31}$$

where $p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{b})$ is the probability mass function of $\boldsymbol{y}$, given $\boldsymbol{\beta}$ and $\boldsymbol{b}$, and $f(\boldsymbol{b}|\boldsymbol{\Sigma})$ is the (Gaussian) probability density at $\boldsymbol{b}$.

Unfortunately the integral in (31) does not have a closed-form solution when $p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{b})$ is binomial. However, we can approximate this integral quite accurately using a *Laplace* approximation. For given values of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ we determine the *conditional modes* of the random effects

$$\tilde{\boldsymbol{b}}(\boldsymbol{\beta},\boldsymbol{\theta}) = \arg\max_{\boldsymbol{b}} p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{b}) f(\boldsymbol{b}|\boldsymbol{\Sigma}(\boldsymbol{\theta})), \tag{32}$$

which are the values of the random effects that maximize the conditional density of the random effects given the data and the model parameters. The conditional modes can be determined by a penalized iteratively reweighted least squares algorithm (PIRLS, see §9.3) where the contribution of the fixed effects parameters, $\boldsymbol{\beta}$, is incorporated as an offset, $\boldsymbol{X}\boldsymbol{\beta}$, and the contribution of the variance components, $\boldsymbol{\theta}$, is incorporated as a penalty term in the weighted least squares fit.

At the conditional modes, $\tilde{\boldsymbol{b}}$, we evaluate the second order Taylor series approximation to the log of the integrand (i.e. the log of the conditional density of $\boldsymbol{b}$) and use its integral as an approximation to the likelihood.

It is the Laplace approximation to the likelihood that is optimized to obtain approximate values of the mle's for the parameters and the corresponding conditional modes of the random effects vector $\boldsymbol{b}$.

9.3 Details of the PIRLS algorithm

Recall from (32) that the conditional modes of the random effects $\tilde{\boldsymbol{b}}(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})$ maximize the conditional density of $\boldsymbol{b}$ given the data and values of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. The penalized iteratively reweighted least squares (PIRLS) algorithm for determining these conditional modes combines characteristic of the iteratively reweighted least squares (IRLS) algorithm for generalized linear models (McCullagh and Nelder, 1989, §2.5) and the penalized least squares representation of a linear mixed model (**?**).

At the rth iteration of the IRLS algorithm the current value of the vector of random effects. $\boldsymbol{b}^{(r)}$ (we use parenthesized superscripts to denote the iteration) produces a linear predictor

$$\boldsymbol{\eta}^{(r)} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}^{(r)} \tag{33}$$

with corresponding mean vector $\boldsymbol{\mu}^{(r)} = \boldsymbol{g}^{-1}\boldsymbol{\eta}^{(r)}$. (The vector-valued link and inverse link functions, $\boldsymbol{g}$ and $\boldsymbol{g}^{-1}$, apply the scalar link and inverse link, g and g^{-1}, componentwise.) A vector of weights and a vector of derivatives of

the form $d\eta/d\mu$ are also evaluated. For convenience of notation we express these as diagonal matrices, $\boldsymbol{W}^{(r)}$ and $\boldsymbol{G}^{(r)}$, although calculations involving these quantities are performed component-wise and not as matrices.

The adjusted dependent variate at iteration r is

$$\boldsymbol{z}^{(r)} = \boldsymbol{\eta}^{(r)} + \boldsymbol{G}^{(r)}\left(\boldsymbol{y} - \boldsymbol{\mu}^{(r)}\right) \tag{34}$$

from which the updated parameter, $\boldsymbol{b}^{(r+1)}$, is determined as the solution to

$$\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{W}^{(r)}\boldsymbol{Z}\boldsymbol{b}^{(r+1)} = \boldsymbol{Z}^{\mathsf{T}}\boldsymbol{W}^{(r)}\boldsymbol{z}^{(r)}. \tag{35}$$

McCullagh and Nelder (1989, §2.5) show that the IRLS algorithm is equivalent to the Fisher scoring algorithm for any link function and also equivalent to the Newton-Raphson algorithm when the link function is the natural link for a probability distribution in the exponential family. That is, IRLS will minimize $-\log p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{b})$ for fixed $\boldsymbol{\beta}$. However, we wish to determine

$$\begin{aligned}\tilde{\boldsymbol{b}}(\boldsymbol{\beta},\boldsymbol{\theta}) &= \arg\max_{\boldsymbol{b}} p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{b}) f(\boldsymbol{b}|\boldsymbol{\Sigma}(\boldsymbol{\theta})) \\ &= \arg\min_{\boldsymbol{b}} \left[-\log p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{b}) + \frac{\boldsymbol{b}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\boldsymbol{b}}{2}\right].\end{aligned} \tag{36}$$

As shown in Bates and DebRoy (2004) we can incorporate the contribution of the Gaussian distribution by adding q "pseudo-observations" with constant unit weights, observed values of 0 and predicted values of $\boldsymbol{\Delta}(\boldsymbol{\theta})\boldsymbol{b}$ where $\boldsymbol{\Delta}$ is any $q \times q$ matrix such that $\boldsymbol{\Delta}^{\mathsf{T}}\boldsymbol{\Delta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})$.

Thus the update in the penalized iteratively reweighted least squares (PIRLS) algorithm for determining the conditional modes, $\tilde{\boldsymbol{b}}(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})$, expresses $\boldsymbol{b}^{(r+1)}$ as the solution to the penalized weighted least squares problem

$$\left(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{W}^{(r)}\boldsymbol{Z} + \boldsymbol{\Sigma}^{-1}\right)\boldsymbol{b}^{(r+1)} = \boldsymbol{Z}^{\mathsf{T}}\boldsymbol{W}^{(r)}\boldsymbol{z}^{(r)}. \tag{37}$$

or the equivalent problem

$$\left(\boldsymbol{Z}^{*\mathsf{T}}\boldsymbol{W}^{(r)}\boldsymbol{Z}^{*} + \boldsymbol{I}\right)\boldsymbol{b}^{*(r+1)} = \boldsymbol{Z}^{*\mathsf{T}}\boldsymbol{W}^{(r)}\boldsymbol{z}^{(r)}. \tag{38}$$

The sequence of iterates $\boldsymbol{b}^{*(0)}, \boldsymbol{b}^{*(1)}, \ldots$ is considered to have converged to the conditional modes $\tilde{\boldsymbol{b}}^{*}(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})$ when the relative change in the linear predictors $\|\boldsymbol{\eta}^{(r+1)} - \boldsymbol{\eta}^{(r)}\|/\|\boldsymbol{\eta}^{(r)}\|$ falls below a threshold. The variance-covariance matrix of $\boldsymbol{b}^{*}$, conditional on $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, is approximated as

$$\mathrm{Var}\left(\boldsymbol{b}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y}\right) \approx \boldsymbol{D} \equiv \left(\boldsymbol{Z}^{*\mathsf{T}}\boldsymbol{W}^{(r)}\boldsymbol{Z}^{*} + \boldsymbol{I}\right)^{-1}. \tag{39}$$

This approximation is analogous to using the inverse of Fisher's information matrix as the approximate variance-covariance matrix for maximum likelihood estimates.

9.4 Details of the Laplace approximation

The Laplace approximation to the likelihood $L(\boldsymbol{\beta},\boldsymbol{\theta}|\boldsymbol{y})$ is obtained by replacing the logarithm of the integrand in (31) by its second-order Taylor series at the conditional maximum, $\tilde{\boldsymbol{b}}(\boldsymbol{\beta},\boldsymbol{\theta})$. On the scale of the deviance (negative twice the log-likelihood) the approximation is

$$\begin{aligned}
-2\ell(\boldsymbol{\beta},\boldsymbol{\theta}|\boldsymbol{y}) &= -2\log\left\{\int_{\boldsymbol{b}} p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{b})f(\boldsymbol{b}|\boldsymbol{\Sigma}(\boldsymbol{\theta}))\,d\boldsymbol{b}\right\}\\
&\approx 2\log\left\{\int_{\boldsymbol{b}}\exp\left\{-\frac{1}{2}\left[d(\boldsymbol{\beta},\tilde{\boldsymbol{b}},\boldsymbol{y})+\tilde{\boldsymbol{b}}^{\mathsf{T}}\tilde{\boldsymbol{b}}^{*}++\boldsymbol{b}^{\mathsf{T}}\boldsymbol{D}^{-1}\boldsymbol{b}\right]\right\}\,d\boldsymbol{b}\right\}\\
&= d(\boldsymbol{\beta},\tilde{\boldsymbol{b}},\boldsymbol{y})+\tilde{\boldsymbol{b}}^{*\mathsf{T}}\tilde{\boldsymbol{b}}^{*}+\log|\boldsymbol{D}|
\end{aligned} \tag{40}$$

where $d(\boldsymbol{\beta},\boldsymbol{b},\boldsymbol{y})$ is the deviance function from the linear predictor only. That is, $d(\boldsymbol{\beta},\boldsymbol{b},\boldsymbol{y}) = -2\log p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{b})$. This quantity can be evaluated as the sum of the deviance residuals (McCullagh and Nelder, 1989, §2.4.3).

References

Douglas M. Bates and Saikat DebRoy. Linear mixed models and penalized least squares. *J. of Multivariate Analysis*, 2004. to appear.

Tim Davis. CHOLMOD: sparse supernodal Cholesky factorization and update/downdate. http://www.cise.ufl.edu/research/sparse/cholmod, 2005a.

Tim Davis. CSparse: a concise sparse matrix package. http://www.cise.ufl.edu/research/sparse/CSparse, 2005b.

Tim Davis. An approximate minimal degree ordering algorithm. *SIAM J. Matrix Analysis and Applications*, 17(4):886–905, 1996.

Timothy A. Davis. *Direct Methods for Sparse Linear Systems*. Fundamentals of Algorithms. SIAM, 2006.

Peter McCullagh and John Nelder. *Generalized Linear Models*. Chapman and Hall, 2nd edition, 1989.

José C. Pinheiro and Douglas M. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer, 2000. ISBN 0-387-98957-9.

A Notation

A.1 Random variables

- Y- the n-dimensional random variable of responses. The observed responses are the n-vector $\boldsymbol{y}$.
- B - The q-dimensional vector of random effects. This vector is not observed directly. It has the properties $\mathsf{E}[B] = \mathbf{0}$ and $\mathsf{Var}([B]) = \sigma^2\boldsymbol{\Sigma}(\boldsymbol{\theta})$, where the scalar σ is the common scale factor (if used in the model) and $\boldsymbol{\Sigma}$ is a $q \times q$ symmetric, positive semi-definite *relative variance-covariance matrix* determined by the variance parameter vector $\boldsymbol{\theta}$.
- U - a q-dimensional *unit* vector of random effects with distribution $U \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I}_q)$.

A.2 Dimensions