Douglas M. Bates

lme4: Mixed-effects modeling with R

January 29, 2010

Springer

Contents

List of Figures

Acronyms

AIC	Akaike's Information Criterion
BIC	Bayesian Information Criterion (also called "Schwatz's Bayesian Information Criterion")
BLUP	Best Linear Unbiased Predictor
LRT	Likelihood Ratio Test
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimate or Maximum Likelihood Estimator
REML	The REstricted (or "REsidual") Maximum Likelihood estimation criterion
PRSS	Penalized, Residual Sum-of-Squares
PWRSS	Penalized, Weighted, Residual Sum-of-Squares

Chapter 1
A Simple, Linear, Mixed-effects Model

In this book we describe the theory behind a type of statistical model called *mixed-effects* models and the practice of fitting and analyzing such models using the `lme4` package for R. These models are used in many different disciplines. Because the descriptions of the models can vary markedly between disciplines, we begin by describing what mixed-effects models are and by exploring a very simple example of one type of mixed model, the *linear mixed model*.

This simple example allows us to illustrate the use of the `lmer` function in the `lme4` package for fitting such models and analyzing the fitted model. Building from the example we describe the general form of linear mixed models that can be fit using `lmer`.

1.1 Mixed-effects Models

Mixed-effects models, like many other types of statistical models, describe a relationship between a *response* variable and some of the *covariates* that have been measured or observed along with the response. In mixed-effects models at least one of the covariates is a *categorical* covariate representing experimental or observational "units" in the data set. In the example from the chemical industry that is given in this chapter, the observational unit is the batch of an intermediate product used in production of a dye. In medical and social sciences the observational units are often the human or animal subjects in the study. In agriculture the experimental units may be the plots of land or the specific plants being studied.

In all of these cases the categorical covariate or covariates are observed at a set of discrete *levels*. We may use numbers, such as subject identifiers, to designate the particular levels that we observed but these numbers are simply labels. The important characteristic of a categorical covariate is that, at each

observed value of the response, the covariate takes on the value of one of a set of distinct levels.

Parameters associated with the particular levels of a covariate are sometimes called the "effects" of the levels. If the set of possible levels of the covariate is fixed and reproducible we model the covariate using *fixed-effects* parameters. If the levels that we observed represent a random sample from the set of all possible levels we incorporate *random effects* in the model.

There are two things to notice about this distinction between fixed-effects parameters and random effects. First, the names are misleading because the distinction between fixed and random is more a property of the levels of the categorical covariate than a property of the effects associated with them. Secondly, we distinguish between "fixed-effects parameters", which are indeed parameters in the statistical model, and "random effects", which, strictly speaking, are not parameters. As we will see shortly, random effects are unobserved random variables.

To make the distinction more concrete, suppose that we wish to model the annual reading test scores for students in a school district and that the covariates recorded with the score include a student identifier and the student's gender. Both of these are categorical covariates. The levels of the gender covariate, male and female, are fixed. If we consider data from another school district or we incorporate scores from earlier tests, we will not change those levels. On the other hand, the students whose scores we observed would generally be regarded as a sample from the set of all possible students whom we could have observed. Adding more data, either from more school districts or from results on previous or subsequent tests, will increase the number of distinct levels of the student identifier.

Mixed-effects models or, more simply, *mixed models* are statistical models that incorporate both fixed-effects parameters and random effects. Because of the way that we will define random effects, a model with random effects always includes at least one fixed-effects parameter. Thus, any model with random effects is a mixed model.

We characterize the statistical model in terms of two random variables: a q-dimensional vector of random effects represented by the random variable $\mathscr{B}$ and an n-dimensional response vector represented by the random variable $\mathscr{Y}$. We observe the value, $\mathbf{y}$, of $\mathscr{Y}$. We do not observe the value of $\mathscr{B}$.

When formulating the model we describe the unconditional distribution of $\mathscr{B}$ and the conditional distribution, $(\mathscr{Y}|\mathscr{B} = \mathbf{b})$. The descriptions of the distributions involve the form of the distribution and the values of certain parameters. We use the observed values of the response and the covariates to estimate these parameters and to make inferences about them.

That the big picture. Now let's make this more concrete by describing a particular, versatile class of mixed models called linear mixed models and by studying a simple example of such a model. First we will describe the data in the example.

1.2 The Dyestuff and Dyestuff2 Data

Models with random effects have been in use for a long time. The first edition of the classic book, *Statistical Methods in Research and Production*, edited by O.L. Davies, was published in 1947 and contained examples of the use of random effects to characterize batch-to-batch variability in chemical processes. The data from one of these examples are available as the `Dyestuff` data in the `lme4` package. In this section we describe and plot these data and introduce a second example, the `Dyestuff2` data, described in Box and Tiao [1973].

1.2.1 The Dyestuff *Data*

The `Dyestuff` data are described in Davies and Goldsmith [1972, Table 6.3, p. 131], the fourth edition of the book mentioned above, as coming from

> an investigation to find out how much the variation from batch to batch in the quality of an intermediate product (H-acid) contributes to the variation in the yield of the dyestuff (Naphthalene Black 12B) made from it. In the experiment six samples of the intermediate, representing different batches of works manufacture, were obtained, and five preparations of the dyestuff were made in the laboratory from each sample. The equivalent yield of each preparation as grams of standard colour was determined by dye-trial.

To access these data within R we must first attach the `lme4` package to our session using

```
> library(lme4)
```

Note that the `">"` symbol in the line shown is the prompt in R and not part of what the user types. The `lme4` package must be attached before any of the data sets or functions in the package can be used. If typing this line results in an error report stating that there is no package by this name then you must first install the package.

In what follows, we will assume that the `lme4` package has been installed and that it has been attached to the R session before any of the code shown has been run.

The `str` function in R provides a concise description of the structure of the data

```
> str(Dyestuff)
```

```
'data.frame':        30 obs. of  2 variables:
 $ Batch: Factor w/ 6 levels "A","B","C","D",..: 1 1 1 1 1 2 2 2 2 2 ...
 $ Yield: num  1545 1440 1440 1520 1580 ...
```

from which we see that it consists of 30 observations of the `Yield`, the response variable, and of the covariate, `Batch`, which is a categorical variable stored as a `factor` object. If the labels for the factor levels are arbitrary, as they are

here, we will use letters instead of numbers for the labels. That is, we label the batches as `"A"` through `"F"` rather than `"1"` through `"6"`. When the labels are letters it is clear that the variable is categorical. When the labels are numbers a categorical covariate can be mistaken for a numeric covariate, with unintended consequences.

It is a good practice to apply `str` to any data frame the first time you work with it and to check carefully that any categorical variables are indeed represented as factors.

The data in a data frame are viewed as a table with columns corresponding to variables and rows to observations. The functions `head` and `tail` print the first or last few rows (the default value of "few" happens to be 6 but we can specify another value if we so choose)

```
> head(Dyestuff)
```

```
  Batch Yield
1     A  1545
2     A  1440
3     A  1440
4     A  1520
5     A  1580
6     B  1540
```

or we could ask for a `summary` of the data

```
> summary(Dyestuff)
```

```
 Batch     Yield
 A:5   Min.   :1440
 B:5   1st Qu.:1469
 C:5   Median :1530
 D:5   Mean   :1528
 E:5   3rd Qu.:1575
 F:5   Max.   :1635
```

Although the `summary` does show us an important property of the data, namely that there are exactly 5 observations on each batch — a property that we will describe by saying that the data are *balanced* with respect to `Batch` — we usually learn much more about the structure of such data from plots like Fig. 1.1 than we can from numerical summaries.

In Fig. 1.1 we can see that there is considerable variability in yield, even for preparations from the same batch, but there is also noticeable batch-to-batch variability. For example, four of the five preparations from batch F provided lower yields than did any of the preparations from batches C and E.

This plot, and essentially all the other plots in this book, were created using Deepayan Sarkar's `lattice` package for R. In Sarkar [2008] he describes how one would create such a plot. Because this book was created using Sweave [Leisch, 2002], the exact code used to create the plot, as well as the code for all the other figures and calculations in the book, is available on the web site for the book. In section **??** we review some of the principles of lattice

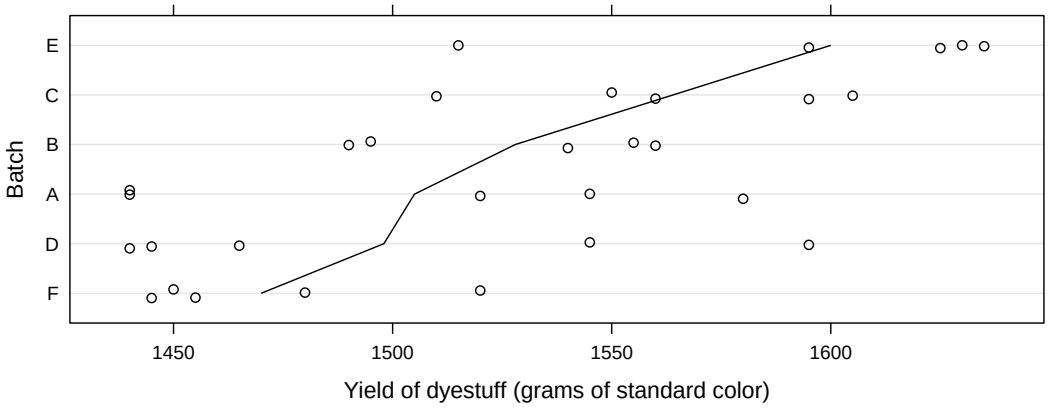


Fig. 1.1 Yield of dyestuff (Napthalene Black 12B) for 5 preparations from each of 6 batches of an intermediate product (H-acid). The line joins the mean yields from the batches, which have been ordered by increasing mean yield. The vertical positions are "jittered" slightly to avoid over-plotting. Notice that the lowest yield for batch A was observed for two distinct preparations from that batch.

graphics, such as reordering the levels of the `Batch` factor by increasing mean response, that enhance the informativeness of the plot. At this point we will concentrate on the information conveyed by the plot and not on how the plot is created.

In section 1.3.1 we will use mixed models to quantify the variability in yield between batches. For the time being let us just note that the particular batches used in this experiment are a selection or sample from the set of all batches that we wish to consider. Furthermore, the extent to which one particular batch tends to increase or decrease the mean yield of the process — in other words, the "effect" of that particular batch on the yield — is not as interesting to us as is the extent of the variability between batches. For the purposes of designing, monitoring and controlling a process we want to predict the yield from future batches, taking into account the batch-to-batch variability and the within-batch variability. Being able to estimate the extent to which a particular batch in the past increased or decreased the yield is not usually an important goal for us. We will model the effects of the batches as random effects rather than as fixed-effects parameters.

1.2.2 The `Dyestuff2` *Data*

The `Dyestuff2` data are simulated data presented in Box and Tiao [1973, Table 5.1.4, p. 247] where the authors state

> These data had to be constructed for although examples of this sort undoubtedly occur in practice they seem to be rarely published.

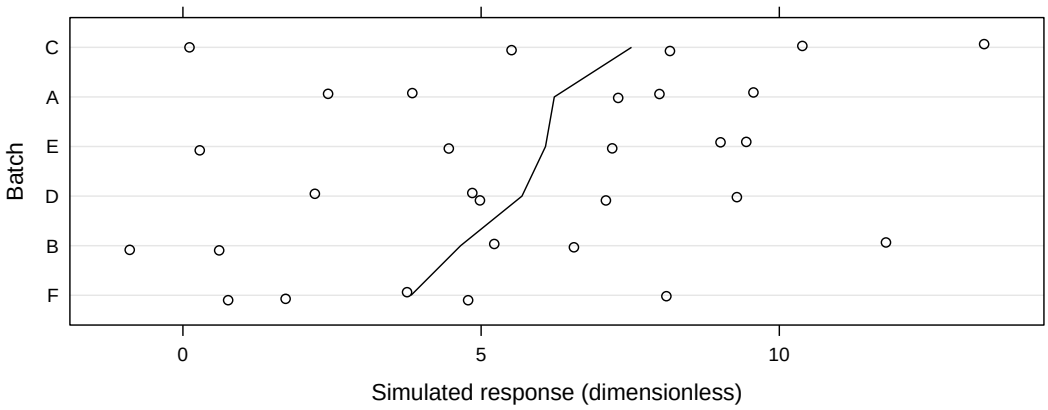


Fig. 1.2 Simulated data presented in Box and Tiao [1973] with a structure similar to that of the `Dyestuff` data. These data represent a case where the batch-to-batch variability is small relative to the within-batch variability.

The structure and summary

```
> str(Dyestuff2)

'data.frame':        30 obs. of  2 variables:
 $ Batch: Factor w/ 6 levels "A","B","C","D",..: 1 1 1 1 1 2 2 2 2 2 ...
 $ Yield: num  7.3 3.85 2.43 9.57 7.99 ...

> summary(Dyestuff2)

 Batch     Yield
 A:5   Min.   :-0.892
 B:5   1st Qu.: 2.765
 C:5   Median : 5.365
 D:5   Mean   : 5.666
 E:5   3rd Qu.: 8.151
 F:5   Max.   :13.434
```

are intentionally similar to those of the `Dyestuff` data. As can be seen in Fig. 1.2 the batch-to-batch variability in these data is small compared to the within-batch variability. In some approaches to mixed models it can be difficult to fit models to such data. Paradoxically, small "variance components" can be more difficult to estimate than large variance components.

The methods we will present are not compromised when estimating small variance components.

1.3 Fitting Linear Mixed Models

Before we formally define a linear mixed model, let's go ahead and fit models to these data sets using `lmer`. Like most model-fitting functions in R, `lmer`

takes, as its first two arguments, a *formula* specifying the model and the *data* with which to evaluate the formula. This second argument, `data`, is optional but recommended. It is usually the name of a data frame, such as those we examined in the last section. Throughout this book all model specifications will be given in this formula/data format.

We will explain the structure of the formula after we have considered an example.

1.3.1 A Model For the `Dyestuff` *Data*

We fit a model to the `Dyestuff` data allowing for an overall level of the `Yield` and for an additive random effect for each level of `Batch`

```
> fm1 <- lmer(Yield ~ 1 + (1|Batch), Dyestuff)
> print(fm1)

Linear mixed model fit by REML
Formula: Yield ~ 1 + (1 | Batch)
   Data: Dyestuff
 REML
319.7

Random effects:
 Groups   Name        Variance Std.Dev.
 Batch    (Intercept) 1764.0   42.001
 Residual             2451.3   49.510
Number of obs: 30, groups: Batch, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)  1527.50      19.38    78.8
```

In the first line we call the `lmer` function to fit a model with formula

```
Yield ~ 1 + (1 | Batch)
```

applied to the `Dyestuff` data and assign the result to the name `fm1`. (The name is arbitrary. I happen to use names that start with `fm`, indicating "fitted model".)

As is customary in R, there is no output shown after this assignment. We have simply saved the fitted model as an object named `fm1`. In the second line we display some information about the fitted model by applying `print` to `fm1`. In later examples we will condense these two steps into one but here it helps to emphasize that we save the result of fitting a model then apply various *extractor* functions to the fitted model to get a brief summary of the model fit or to obtain the values of some of the estimated quantities.

1.3.1.1 Details of the Printed Display

The printed display of a model fit with `lmer` has four major sections: a description of the model that was fit, some statistics characterizing the model fit, a summary of properties of the random effects and a summary of the fixed-effects parameter estimates. We consider each of these sections in turn.

The description section states that this is a linear mixed model in which the parameters have been estimated as those that minimize the REML criterion (explained in section **??**). The `formula` and `data` arguments are displayed for later reference. If other, optional arguments affecting the fit, such as a `subset` specification, were used, they too will be displayed here.

For models fit by the REML criterion the only statistic describing the model fit is the value of the REML criterion itself. An alternative set of parameter estimates, the maximum likelihood estimates, are obtained by specifying the optional argument `REML = FALSE`.

```
> (fm1ML <- lmer(Yield ~ 1 + (1|Batch), Dyestuff, REML = FALSE))

Linear mixed model fit by maximum likelihood
Formula: Yield ~ 1 + (1 | Batch)
   Data: Dyestuff
   AIC   BIC logLik deviance
 333.3 337.5 -163.7    327.3

Random effects:
 Groups   Name        Variance Std.Dev.
 Batch    (Intercept) 1388.3   37.26
 Residual             2451.3   49.51
Number of obs: 30, groups: Batch, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)  1527.50      17.69   86.33
```

(Notice that this code fragment also illustrates a way to condense the assignment and the display of the fitted model into a single step. The redundant set of parentheses surrounding the assignment causes the result of the assignment to be displayed. We will use this device often in what follows.)

The display of a model fit by maximum likelihood provides several other model-fit statistics such as Akaike's Information Criterion (`AIC`) [Sakamoto et al., 1986], Schwarz's Bayesian Information Criterion (`BIC`) [Schwarz, 1978], the log-likelihood (`logLik`) at the parameter estimates, and the deviance (negative twice the log-likelihood) at the parameter estimates. These are all statistics related to the model fit and are used to compare different models fit to the same data.

At this point the important thing to note is that the default estimation criterion is the REML criterion. Generally the REML estimates of variance components are preferred to the ML estimates. However, when comparing

models it is safest to refit all the models using the maximum likelihood criterion. We will discuss comparisons of model fits later in section **??**.

The third section is the table of estimates of parameters associated with the random effects. There are two sources of variability in the model we have fit, a batch-to-batch variability in the level of the response and the residual or per-observation variability — also called the within-batch variability. The name "residual" is used in statistical modeling to denote the part of the variability that cannot be explained or modeled with the other terms. It is the variation in the observed data that is "left over" after we have determined the estimates of the parameters in the other parts of the model.

Some of the variability in the response is associated with the fixed-effects terms. In this model there is only one such term, labeled as the `(Intercept)`. The name "intercept", which is better suited to models based on straight lines written in a slope/intercept form, should be understood to represent an overall "typical" or mean level of the response in this case. (In case you are wondering about the parentheses around the name, they are included so that you can't accidentally create a variable with a name that conflicts with this name.) The line labeled `Batch` in the random effects table shows that the random effects added to the `(Intercept)` term, one for each level of the `Batch` factor, are modeled as random variables whose unconditional variance is estimated as 1764.05 g^2 in the REML fit and as 1388.33 g^2 in the ML fit. The corresponding standard deviations are 42.00 g for the REML fit and 37.26 g for the ML fit.

Note that the last column in the random effects summary table is the estimate of the variability expressed as a standard deviation rather than as a variance. These are provided because it is usually easier to visualize standard deviations, which are on the scale of the response, than it is to visualize the magnitude of a variance. The values in this column are a simple re-expression (the square root) of the estimated variances. Do not confuse them with the standard errors of the variance estimators, which are not given here. In section 1.5 we explain why we do not provide standard errors of variance estimates.

The line labeled `Residual` in this table gives the estimate of the variance of the residuals (also in g^2) and its corresponding standard deviation. For the REML fit the estimated standard deviation of the residuals is 49.51 g and for the ML fit it is also 49.51 g (Generally these estimates do not need to be equal. They happen to be equal in this case because of the simple model form and the balanced data set.)

The last line in the random effects table states the number of observations to which the model was fit and the number of levels of any "grouping factors" for the random effects. In this case we have a single random effects term, `(1|Batch)`, in the model formula and the grouping factor for that term is `Batch`. There will be a total of six random effects, one for each level of `Batch`.

The final part of the printed display gives the estimates and standard errors of any fixed-effects parameters in the model. The only fixed-effects term in

the model formula is the `1`, denoting a constant which, as explained above, is labeled as `(Intercept)`. For both the REML and the ML estimation criterion the estimate of this parameter is 1527.5 g (equality is again a consequence of the simple model and balanced data set). The standard error of the intercept estimate is 19.38 g for the REML fit and 17.69 g for the ML fit.

1.3.2 A Model For the `Dyestuff2` *Data*

Fitting a similar model to the `Dyestuff2` data produces an estimate $\widehat{\sigma}_1 = 0$ in both the REML

```
> (fm2 <- lmer(Yield ~ 1 + (1|Batch), Dyestuff2))

Linear mixed model fit by REML
Formula: Yield ~ 1 + (1 | Batch)
   Data: Dyestuff2
 REML
161.8

Random effects:
 Groups   Name        Variance Std.Dev.
 Batch    (Intercept)  0.000   0.0000
 Residual             13.806   3.7157
Number of obs: 30, groups: Batch, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)   5.6656     0.6784   8.352
```

and the ML fits.

```
> (fm2ML <- update(fm2, REML = FALSE))

Linear mixed model fit by maximum likelihood
Formula: Yield ~ 1 + (1 | Batch)
   Data: Dyestuff2
   AIC   BIC logLik deviance
 168.9 173.1 -81.44    162.9

Random effects:
 Groups   Name        Variance Std.Dev.
 Batch    (Intercept)  0.000   0.0000
 Residual             13.346   3.6532
Number of obs: 30, groups: Batch, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)    5.666      0.667   8.494
```

(Note the use of the `update` function to re-fit a model changing some of the arguments. In a case like this, where the call to fit the original model is not

very complicated, the use of `update` is not that much simpler than repeating the original call to `lmer` with extra arguments. For complicated model fits it can be.)

An estimate of 0 for σ_1 does not mean that there is no variation between the groups. Indeed Fig. 1.2 shows that there is some small amount of variability between the groups. The estimate, $\widehat{\sigma}_1 = 0$, simply indicates that the level of "between-group" variability is not sufficient to warrant incorporating random effects in the model.

The important point to take away from this example is that we must allow for the estimates of variance components to be zero. We describe such a model as being degenerate, in the sense that it corresponds to a linear model in which we have removed the random effects associated with `Batch`. Degenerate models can and do occur in practice. Even when the final fitted model is not degenerate, we must allow for such models when determining the parameter estimates through numerical optimization.

To reiterate, the model `fm2` corresponds to the linear model

```
> summary(fm2a <- lm(Yield ~ 1, Dyestuff2))

Call:
lm(formula = Yield ~ 1, data = Dyestuff2)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5576 -2.9006 -0.3006  2.4854  7.7684

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.6656     0.6784   8.352 3.32e-09

Residual standard error: 3.716 on 29 degrees of freedom
```

because the random effects are inert, in the sense that they have a variance of zero, and can be removed.

Notice that the estimate of σ from the linear model (called the `Residual standard error` in the output corresponds to the estimate in the REML fit (`fm2`) but not that from the ML fit (`fm2ML`). The fact that the REML estimates of variance components generalize the estimate of the variance used in linear models, in the sense that they correspond in the degenerate case, is part of the motivation for the use of the REML criterion for fitting mixed-effects models.

1.3.3 Further Assessment of the Fitted Models

The parameter estimates in a statistical model represent our "best guess" at the unknown values of the model parameters and, as such, are important

results in statistical modeling. However, they are not the whole story. Statistical models characterize the variability in the data and we must assess the effect of this variability on the parameter estimates and on the precision of predictions made from the model.

In section 1.5 we introduce a method of assessing variability in parameter estimates using the "profiled deviance" and in section 1.6 we show methods of characterizing the conditional distribution of the random effects given the data. Before we get to these sections, however, we should state in some detail the probability model for linear mixed-effects and establish some definitions and notation. In particular, before we can discuss profiling the deviance, we should define the deviance. We do that in the next section.

1.4 The Linear Mixed-effects Probability Model

In explaining some of parameter estimates related to the random effects we have used terms such as "unconditional distribution" from the theory of probability. Before proceeding further we should clarify the linear mixed-effects probability model and define several terms and concepts that will be used throughout the book.

1.4.1 Definitions and Results

In this section we provide some definitions and formulas without derivation and with minimal explanation, so that we can use these terms in what follows. In Chapter 4 we revisit these definitions providing derivations and more explanation.

As mentioned in section 1.1, a mixed model incorporates two random variables: $\mathscr{B}$, the q-dimensional vector of random effects, and $\mathscr{Y}$, the n-dimensional response vector. In a linear mixed model the unconditional distribution of $\mathscr{B}$ and the conditional distribution, $(\mathscr{Y}|\mathscr{B}=\mathbf{b})$, are both multivariate Gaussian (or "normal") distributions,

$$\begin{aligned}(\mathscr{Y}|\mathscr{B}=\mathbf{b}) &\sim \mathscr{N}(\mathbf{X}\beta+\mathbf{Z}\mathbf{b},\sigma^2\mathbf{I})\\ \mathscr{B} &\sim \mathscr{N}(\mathbf{0},\Sigma_\theta).\end{aligned} \tag{1.1}$$

The *conditional mean* of $\mathscr{Y}$, given $\mathscr{B}=\mathbf{b}$, is the *linear predictor*, $\mathbf{X}\beta+\mathbf{Z}\mathbf{b}$, which depends on the p-dimensional *fixed-effects parameter*, β, and on $\mathbf{b}$. The *model matrices*, $\mathbf{X}$ and $\mathbf{Z}$, of dimension $n\times p$ and $n\times q$, respectively, are determined from the formula for the model and the values of covariates. Although the matrix $\mathbf{Z}$ can be large (i.e. both n and q can be large), it is sparse (i.e. most of the elements in the matrix are zero).

The *relative covariance factor*, Λ_θ is a $q \times q$ matrix, depending on the *variance-component parameter*, θ, and generating the symmetric $q \times q$ variance-covariance matrix, Σ_θ, according to

$$\Sigma_\theta = \sigma^2 \Lambda_\theta \Lambda_\theta^\mathsf{T}. \tag{1.2}$$

The *spherical random effects*, $\mathscr{U} \sim \mathscr{N}(\mathbf{0}, \sigma^2 \mathbf{I}_q)$, determine $\mathscr{B}$ according to

$$\mathscr{B} = \Lambda_\theta \mathscr{U}.$$

The *penalized residual sum of squares* (PRSS),

$$r^2(\theta, \beta, \mathbf{u}) = \left\{ \|\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\Lambda_\theta \mathbf{u}\|^2 + \|\mathbf{u}\|^2 \right\}, \tag{1.3}$$

is the sum of the residual sum of squares, measuring fidelity of the model to the data, and a penalty on the size of $\mathbf{u}$, measuring the complexity of the model. Minimizing r^2 with respect to $\mathbf{u}$,

$$r^2_{\beta,\theta} = \min_{\mathbf{u}} \left\{ \|\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\Lambda_\theta \mathbf{u}\|^2 + \|\mathbf{u}\|^2 \right\} \tag{1.4}$$

is a direct (i.e. non-iterative) computation for which we calculate the *sparse Cholesky factor*, $\mathbf{L}_\theta$, which is a lower triangular $q \times q$ matrix satisfying

$$\mathbf{L}_\theta \mathbf{L}_\theta^\mathsf{T} = \Lambda_\theta^\mathsf{T} \mathbf{Z}^\mathsf{T} \mathbf{Z} \Lambda_\theta + \mathbf{I}_q. \tag{1.5}$$

where $\mathbf{I}_q$ is the $q \times q$ *identity matrix*.

The *deviance* (negative twice the log-likelihood) of the parameters, given the data, $\mathbf{y}$, is

$$d(\theta, \beta, \sigma | \mathbf{y}) = n \log(2\pi\sigma^2) + \log(|\mathbf{L}_\theta|^2) + \frac{r^2_{\beta,\theta}}{\sigma^2}. \tag{1.6}$$

where $|\mathbf{L}_\theta|$ denotes the *determinant* of $\mathbf{L}_\theta$. Because $\mathbf{L}_\theta$ is triangular, its determinant is the product of its diagonal elements.

Because the conditional mean, μ, is a linear function of β and $\mathbf{u}$, minimization of the PRSS with respect to both β and $\mathbf{u}$ to produce

$$r^2_\theta = \min_{\beta,\mathbf{u}} \left\{ \|\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\Lambda_\theta \mathbf{u}\|^2 + \|\mathbf{u}\|^2 \right\} \tag{1.7}$$

is also a direct calculation. The values of $\mathbf{u}$ and β that provide this minimum are called, respectively, the *conditional mode*, $\tilde{\mathbf{u}}_\theta$, of the spherical random effects and the conditional estimate, $\widehat{\beta}_\theta$, of the fixed effects. At the conditional estimate of the fixed effects the deviance is

$$d(\theta, \widehat{\beta}_\theta, \sigma | \mathbf{y}) = n \log(2\pi\sigma^2) + \log(|\mathbf{L}_\theta|^2) + \frac{r^2_\theta}{\sigma^2}. \tag{1.8}$$

Minimizing this expression with respect to σ^2 produces the conditional estimate

$$\widehat{\sigma^2}_\theta = \frac{r_\theta^2}{n} \tag{1.9}$$

which provides the *profiled deviance*,

$$\tilde{d}(\theta|\mathbf{y}) = d(\theta, \widehat{\beta}_\theta, \widehat{\sigma}_\theta|\mathbf{y}) = \log(|\mathbf{L}_\theta|^2) + n\left(1 + \log\left(\frac{2\pi r_\theta^2}{n}\right)\right), \tag{1.10}$$

a function of θ alone.

The *maximum likelihood estimate* (MLE) of θ, written $\widehat{\theta}$, is the value that minimizes the profiled deviance (1.10). We determine this value by numerical optimization. In the process of evaluating $\tilde{d}(\widehat{\theta}|\mathbf{y})$ we determine $\widehat{\beta}$, $\tilde{\mathbf{u}}_{\widehat{\theta}}$ and $r^2_{\widehat{\theta}}$, from which we can evaluate $\widehat{\sigma} = \sqrt{r^2_{\widehat{\theta}}/n}$.

The elements of the conditional mode of $\mathscr{B}$, evaluated at the parameter estimates,

$$\tilde{b}_{\widehat{\theta}} = \Lambda_{\widehat{\theta}} \tilde{u}_{\widehat{\theta}} \tag{1.11}$$

are sometimes called the *best linear unbiased predictors* or BLUPs of the random effects. Although it has an appealing acronym, I don't find the term particularly instructive (what is a "linear unbiased predictor" and in what sense are these the "best"?) and prefer the term "conditional mode", which is explained in section 1.6.

1.4.2 Matrices and Vectors in the Fitted Model Object

The optional argument, `verbose = TRUE`, in a call to `lmer` produces output showing the progress of the iterative optimization of $\tilde{d}(\theta|\mathbf{y})$.

```
> fm1ML <- lmer(Yield ~ 1|Batch, Dyestuff, REML = FALSE, verbose = TRUE)

  0:     327.76702:  1.00000
  1:     327.35312: 0.807151
  2:     327.33414: 0.725317
  3:     327.32711: 0.754925
  4:     327.32706: 0.752678
  5:     327.32706: 0.752578
  6:     327.32706: 0.752581
```

The algorithm converges in 6 iterations to a profiled deviance of 327.32706 at θ =0.752581.

The actual values of many of the matrices and vectors defined above are available in the *environment* of the fitted model object, accessed with the `env` function. For example, $\Lambda_{\widehat{\theta}}$ is

```
> env(fm1ML)$Lambda
```

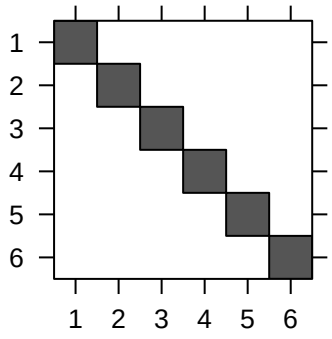


Fig. 1.3 Image of the relative covariance factor, $\Lambda_{\widehat{\theta}}$ for model `fm1ML`. The non-zero elements are shown as darkened squares. The zero elements are blank.

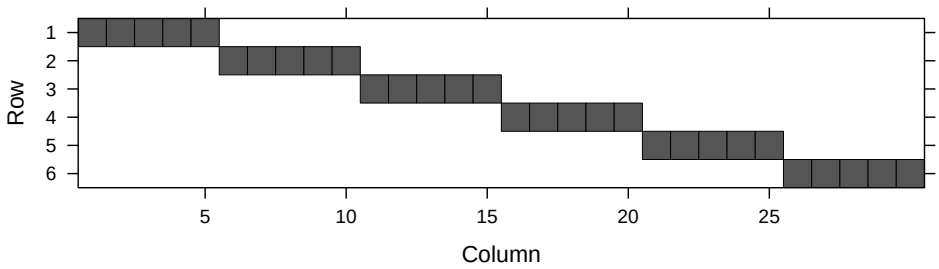


Fig. 1.4 Image of the transpose of the random-effects model matrix, $\mathbf{Z}$, for model `fm1`. The non-zero elements, which are all unity, are shown as darkened squares. The zero elements are blank.

```
6 x 6 diagonal matrix of class "ddiMatrix"
     [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
[1,] 0.7525806         .         .         .         .         .
[2,]         . 0.7525806         .         .         .         .
[3,]         .         . 0.7525806         .         .         .
[4,]         .         .         . 0.7525806         .         .
[5,]         .         .         .         . 0.7525806         .
[6,]         .         .         .         .         . 0.7525806
```

Often we will show the structure of sparse matrices as an image (Fig. 1.3). Especially for large sparse matrices, the image conveys the structure more compactly than does the printed representation.

In this simple model $\Lambda = \theta\mathbf{I}_6$ is a multiple of the identity matrix and the 30×6 model matrix $\mathbf{Z}$, whose transpose is shown in Fig. 1.4, consists of the indicator columns for `Batch`. Because the data are balanced with respect to `Batch`, the Cholesky factor, $\mathbf{L}$ is also a multiple of the identity (you can check this with `image(env(fm1ML)$L)`. The vectors $\mathbf{u}$ and $\mathbf{b}$ and the matrix $\mathbf{X}$ have the same names in `env(fm1ML)`. The vector β is called `fixef`.

1.5 Assessing the Variability of the Parameter Estimates

In this section we show how to create a *profile deviance* object from a fitted linear mixed model and how to use this object to evaluate confidence intervals on the parameters. We also discuss the construction and interpretation of *profile zeta* plots for the parameters and *profile pairs* plots for parameter pairs.

1.5.1 Confidence Intervals on the Parameters

The mixed-effects model fit as `fm1` or `fm1ML` has three parameters for which we obtained estimates. These parameters are σ_1, the standard deviation of the random effects, σ, the standard deviation of the residual or "per-observation" noise term and β_0, the fixed-effects parameter that is labeled as `(Intercept)`.

The `profile` function systematically varies the parameters in a model, assessing the best possible fit that can be obtained with one parameter fixed at a specific value and comparing this fit to the *globally optimal fit*, which is the original model fit that allowed all the parameters to vary. The models are compared according to the change in the deviance, which is the *likelihood ratio test* (LRT) statistic. We apply a *signed square root* transformation to this statistic and plot the resulting function, called ζ, versus the parameter value. A ζ value can be compared to the quantiles of the *standard normal distribution*, $\mathcal{Z} \sim \mathcal{N}(0,1)$. For example, a 95% profile deviance confidence interval on the parameter consists of the values for which $-1.960 < \zeta < 1.960$.

Because the process of profiling a fitted model, which involves re-fitting the model many times, can be computationally intensive, one should exercise caution with complex models fit to very large data sets. Because the statistic of interest is a likelihood ratio, the model is re-fit according to the maximum likelihood criterion, even if the original fit is a REML fit. Thus, there is a slight advantage in starting with an ML fit.

```
> pr1 <- profile(fm1ML)
```

Plots of ζ versus the parameter being profiled (Fig. 1.5) are obtained with

```
> xyplot(pr1, aspect = 1.3)
```

We will refer to such plots as *profile zeta* plots. I usually adjust the aspect ratio of the panels in profile zeta plots to, say, `aspect = 1.3` and frequently set the layout so the panels form a single row (`layout = c(3,1)`, in this case).

The vertical lines in the panels delimit the 50%, 80%, 90%, 95% and 99% confidence intervals, when these intervals can be calculated. Numerical values of the endpoints are returned by the `confint` extractor.

```
> confint(pr1)
```

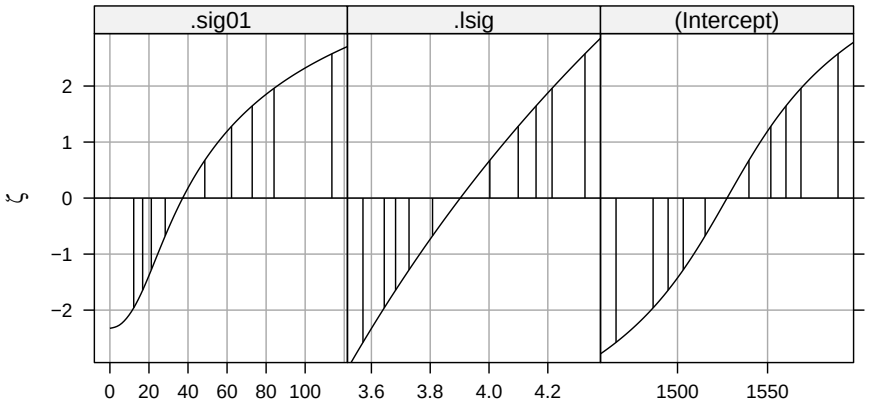


Fig. 1.5 Signed square root, ζ, of the likelihood ratio test statistic for each of the parameters in model `fm1ML`. The vertical lines are the endpoints of 50%, 80%, 90%, 95% and 99% confidence intervals derived from this test statistic.

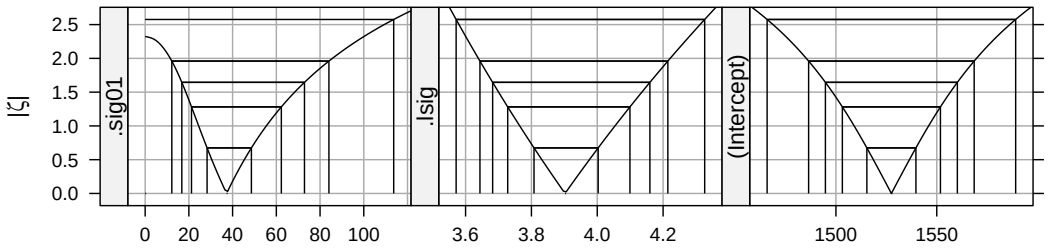


Fig. 1.6 Profiled deviance, on the scale $|\zeta|$, the square root of the change in the deviance, for each of the parameters in model `fm1ML`. The intervals shown are 50%, 80%, 90%, 95% and 99% confidence intervals based on the profile likelihood.

```
                  2.5 %      97.5 %
.sig01        12.197461   84.063361
.lsig          3.643624    4.214461
(Intercept) 1486.451506 1568.548494
```

By default the 95% confidence interval is returned. The optional argument, `level`, is used to obtain other confidence levels.

```
> confint(pr1, level = 0.99)
```

```
                  0.5 %      99.5 %
.sig01               NA  113.690280
.lsig          3.571290    4.326337
(Intercept) 1465.872875 1589.127125
```

Notice that the lower bound on the 99% confidence interval for σ_1 is not defined. Also notice that we profile $\log(\sigma)$ instead of σ, the residual standard deviation.

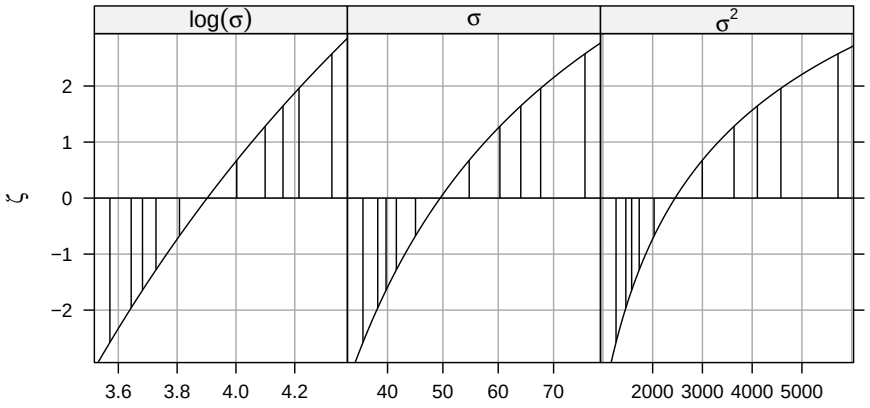


Fig. 1.7 Signed square root, ζ, of the likelihood ratio test statistic as a function of $\log(\sigma)$, of σ and of σ^2. The vertical lines are the endpoints of 50%, 80%, 90%, 95% and 99% confidence intervals.

A plot of $|\zeta|$, the absolute value of ζ, versus the parameter (Fig. 1.6), obtained by adding the optional argument `absVal = TRUE` to the call to `xyplot`, can be more effective for visualizing the confidence intervals.

1.5.2 Interpreting the Profile Zeta Plot

A profile zeta plot, such as Fig. 1.5, shows us the sensitivity of the model fit to changes in the value of particular parameters. Although this is not quite the same as describing the distribution of an estimator, it is a similar idea and we will use some of the terminology from distributions when describing these plots. Essentially we view the patterns in the plots as we would those in a normal probability plot of data values or residuals from a model.

Ideally the profile zeta plot will be close to a straight line over the region of interest, in which case we can perform reliable statistical inference based on the parameter's estimate, its standard error and quantiles of the standard normal distribution. We will describe such as situation as providing a good normal approximation for inference. The common practice of quoting a parameter estimate and its standard error assumes that this is always the case.

In Fig. 1.5 the profile zeta plot for $\log(\sigma)$ is reasonably straight so $\log(\sigma)$ has a good normal approximation. But this does not mean that there is a good normal approximation for σ^2 or even for σ. As shown in Fig. 1.7 the profile zeta plot for $\log(\sigma)$ is slightly skewed, that for σ is moderately skewed and the profile zeta plot for σ^2 is highly skewed. Deviance-based confidence intervals on σ^2 are quite asymmetric, of the form "estimate minus a little, plus a lot".

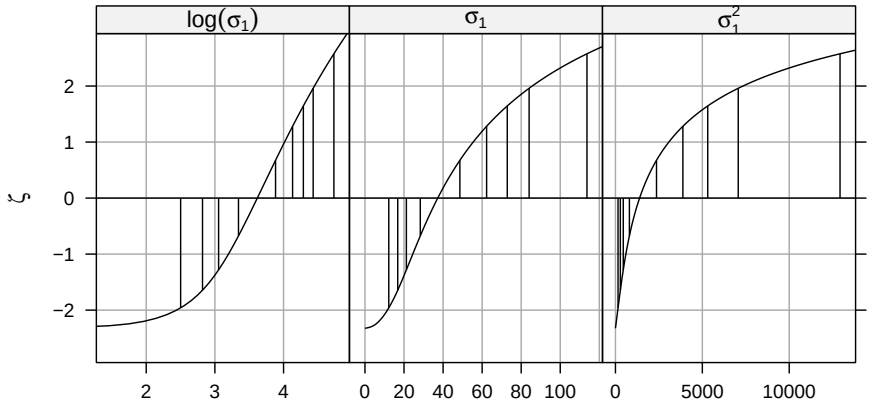


Fig. 1.8 Signed square root, ζ, of the likelihood ratio test statistic as a function of $\log(\sigma_1)$, of σ_1 and of σ_1^2. The vertical lines are the endpoints of 50%, 80%, 90%, 95% and 99% confidence intervals.

This should not come as a surprise to anyone who learned in an introductory statistics course that, given a random sample of data assumed to come from a Gaussian distribution, we use a χ^2 distribution, which can be quite skewed, to form a confidence interval on σ^2. Yet somehow there is a widespread belief that the distribution of variance estimators in much more complex situations should be well approximated by a normal distribution. It is nonsensical to believe that. In most cases summarizing the precision of a variance component estimate by giving an approximate standard error is woefully inadequate.

The pattern in the profile plot for β_0 is sigmoidal (i.e. an elongated "S"-shape). The pattern is symmetric about the estimate but curved in such a way that the profile-based confidence intervals are wider than those based on a normal approximation. We characterize this pattern as symmetric but over-dispersed (relative to a normal distribution). Again, this pattern is not unexpected. Estimators of the coefficients in a linear model without random effects have a distribution which is a scaled Student's T distribution. That is, they follow a symmetric distribution that is over-dispersed relative to the normal.

The pattern in the profile zeta plot for σ_1 is more complex. Fig. 1.8 shows the profile zeta plot on the scale of $\log(\sigma_1)$, σ_1 and σ_1^2. Notice that the profile zeta plot for $\log(\sigma_1)$ is very close to linear to the right of the estimate but flattens out on the left. That is, σ_1 behaves like σ in that its profile zeta plot is more-or-less a straight line on the logarithmic scale, except when σ_1 is close to zero. The model loses sensitivity to values of σ_1 that are close to zero. If, as in this case, zero is within the "region of interest" then we should expect that the profile zeta plot will flatten out on the left hand side.

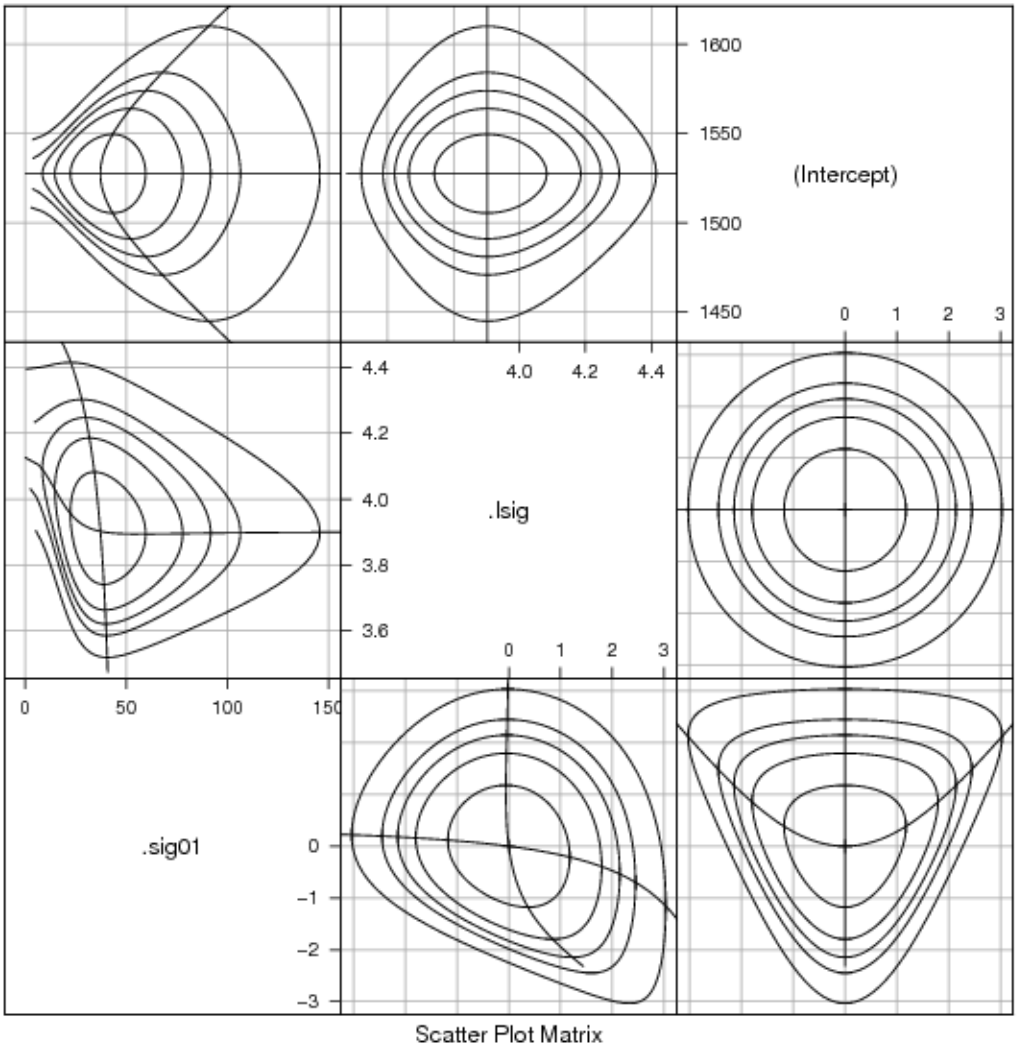


Fig. 1.9 Profile pairs plot for the parameters in model `fm1`. The contour lines correspond to two-dimensional 50%, 80%, 90%, 95% and 99% marginal confidence regions based on the likelihood ratio. Panels below the diagonal represent the (ζ_i, ζ_j) parameters; those above the diagonal represent the original parameters.

1.5.3 Profile Pairs Plots

A profiled deviance object, such as `pr1`, not only provides information on the sensitivity of the model fit to changes in parameters, it also tells us how the parameters influence each other. When we re-fit the model subject to a constraint such as, say, $\sigma_1 = 60$, we obtain the conditional estimates for the other parameters — σ and β_0 in this case. The conditional estimate of, say, σ as a function of σ_1 is called the *profile trace* of σ on σ_1. Plotting such traces provides valuable information on how the parameters in the model are influenced by each other.

The *profile pairs* plot, obtained as

```
> splom(pr1)
```

and shown in Fig. 1.9 shows the profile traces along with interpolated contours of the two-dimensional profiled deviance function. The contours are chosen to correspond to the two-dimensional marginal confidence regions at particular confidence levels.

Because this plot may be rather confusing at first we will explain what is shown in each panel. To make it easier to refer to panels we assign them (x,y) coordinates, as in a Cartesian coordinate system. The columns are numbered 1 to 3 from left to right and the rows are numbered 1 to 3 from bottom to top. Note that the rows are numbered from the bottom to the top, like the y-axis of a graph, not from top to bottom, like a matrix.

The diagonal panels show the ordering of the parameters: σ_1 first, then $\log(\sigma)$ then β_0. Panels above the diagonal are in the original scale of the parameters. That is, the top-left panel, which is the $(1,3)$ position, has σ_1 on the horizontal axis and β_0 on the vertical axis.

In addition to the contour lines in this panel, there are two other lines, which are the profile traces of σ_1 on β_0 and of β_0 on σ_1. The profile trace of β_0 on σ_1 is a straight horizontal line, indicating that the conditional estimate of β_0, given a value of σ_1, is constant. Again, this is a consequence of the simple model form and the balanced data set. The other line in this panel, which is the profile trace of σ_1 on β_0, is curved. That is, the conditional estimate of σ_1 given β_0 depends on β_0. As β_0 moves away from the estimate, $\widehat{\beta}_0$, in either direction, the conditional estimate of σ_1 increases.

We will refer to the two traces on a panel as the "horizontal trace" and "vertical trace". They are not always perfectly horizontal and vertical lines but the meaning should be clear from the panel because one trace will always be more horizontal and the other will be more vertical. The one that is more horizontal is the trace of the parameter on the y axis as a function of the parameter on the horizontal axis, and vice versa.

The contours shown on the panel are interpolated from the profile zeta function and the profile traces, in the manner described in Bates and Watts [1988, Chapter 6]. One characteristic of a profile trace, which we can verify visually in this panel, is that the tangent to a contour must be vertical where it intersects the horizontal trace and horizontal where it intersects the vertical trace.

The $(2,3)$ panel shows β_0 versus $\log(\sigma)$. In this case the traces actually are horizontal and vertical straight lines. That is, the conditional estimate of β_0 doesn't depend on $\log(\sigma)$ and the conditional estimate of $\log(\sigma)$ doesn't depend on β_0. Even in this case, however, the contour lines are not concentric ellipses, because the deviance is not perfectly quadratic in these parameters. That is, the zeta functions, $\zeta(\beta_0)$ and $\zeta(\log(\sigma))$, are not linear.

The $(1,2)$ panel, showing $\log(\sigma)$ versus σ_1 shows distortion along both axes and nonlinear patterns in both traces. When σ_1 is close to zero the conditional estimate of $\log(\sigma)$ is larger than when σ_1 is large. In other words

small values of σ_1 inflate the estimate of $\log(\sigma)$ because the variability that would be explained by the random effects gets incorporated into the residual noise term.

Panels below the diagonal are on the ζ scale, which is why the axes on each of these panels span the same range, approximately -3 to $+3$, and the profile traces always cross at the origin. Thus the $(3,1)$ panel shows $\zeta(\sigma_1)$ on the vertical axis versus $\zeta(\beta_0)$ on the horizontal. These panels allow us to see distortions from an elliptical shape due to nonlinearity of the traces, separately from the one-dimensional distortions caused by a poor choice of scale for the parameter. The ζ scales provide, in some sense, the best possible set of single-parameter transformations for assessing the contours. On the ζ scales the extent of a contour on the horizontal axis is exactly the same as the extent on the vertical axis and both are centered about zero.

Another way to think of this is that, if we would have profiled σ_1^2 instead of σ_1, we would change all the panels in the first column but the panels on the first row would remain the same.

1.6 Assessing the Random Effects

In section 1.4.1 we mentioned that what are sometimes called the BLUPs (or best linear unbiased estimators) of the random effects, $\mathscr{B}$, are the conditional modes evaluated at the parameter estimates, and that they can be calculated as $\tilde{b}_{\widehat{\theta}} = \Lambda_{\widehat{\theta}}\tilde{u}_{\widehat{\theta}}$.

These values are often considered as some sort of "estimates" of the random effects. It can be helpful to think of them this way but it can also be misleading. As we have stated, the random effects are not, strictly speaking, parameters—they are unobserved random variables. We don't estimate the random effects in the same sense that we estimate parameters. Instead, we consider the conditional distribution of $\mathscr{B}$ given the observed data, $(\mathscr{B}|\mathscr{Y} = \mathbf{y})$.

Because the unconditional distribution, $\mathscr{B} \sim \mathscr{N}(\mathbf{0}, \Sigma_\theta)$ is continuous, the conditional distribution, $(\mathscr{B}|\mathscr{Y} = \mathbf{y})$ will also be continuous. In general, the mode of a probability density is the point of maximum density, so the phrase "conditional mode" refers to the point at which this conditional density is maximized. Because this definition relates to the probability model, the values of the parameters are assumed to be known. In practice, of course, we don't know the values of the parameters (if we did there would be no purpose in forming the parameter estimates), so we use the estimated values of the parameters to evaluate the conditional modes.

Those who are familiar with the multivariate Gaussian distribution may recognize that, because both $\mathscr{B}$ and $(\mathscr{Y}|\mathscr{B} = \mathbf{b})$ are multivariate Gaussian, $(\mathscr{B}|\mathscr{Y} = \mathbf{y})$ will also be multivariate Gaussian and the conditional mode will also be the conditional mean of $\mathscr{B}$, given $\mathscr{Y} = \mathbf{y}$. This is the case for a linear

mixed model but it does not carry over to other forms of mixed models. In the general case all we can say about $\tilde{\mathbf{u}}$ or $\tilde{\mathbf{b}}$ is that they maximize a conditional density, which is why we use the term "conditional mode" to describe these values. We will only use the term "conditional mean" and the symbol, μ, in reference to $\mathrm{E}(\mathscr{Y}|\mathscr{B}=\mathbf{b})$, which is the conditional mean of $\mathscr{Y}$ given $\mathscr{B}$, and an important part of the formulation of all types of mixed-effects models.

The `ranef` extractor returns the conditional modes.

```
> ranef(fm1ML)

$Batch
  (Intercept)
A  -16.628221
B    0.369516
C   26.974670
D  -21.801445
E   53.579824
F  -42.494343
```

Applying `str` to the result of `ranef`

```
> str(ranef(fm1ML))

List of 1
 $ Batch:'data.frame':        6 obs. of  1 variable:
  ..$ (Intercept): num [1:6] -16.628 0.37 26.975 -21.801 53.58 ...
 - attr(*, "class")= chr "ranef.mer"
```

shows that the value is a list of data frames. In this case the list is of length 1 because there is only one random-effects term, `(1|Batch)`, in the model and, hence, only one grouping factor, `Batch`, for the random effects. There is only one column in this data frame because the random-effects term, `(1|Batch)`, is a simple, scalar term.

To make this more explicit, random-effects terms in the model formula are those that contain the vertical bar (`"|"`) character. The `Batch` variable is the grouping factor for the random effects generated by this term. An expression for the grouping factor, usually just the name of a variable, occurs to the right of the vertical bar. If the expression on the left of the vertical bar is `1`, as it is here, we describe the term as a *simple, scalar, random-effects term*. The designation "scalar" means there will be exactly one random effect generated for each level of the grouping factor. A simple, scalar term generates a block of indicator columns — the indicators for the grouping factor — in $\mathbf{Z}$. Because there is only one random-effects term in this model and because that term is a simple, scalar term, the model matrix $\mathbf{Z}$ for this model is the indicator matrix for the levels of `Batch`.

In the next chapter we fit models with multiple simple, scalar terms and, in subsequent chapters, we extend random-effects terms beyond simple, scalar terms. When we have only simple, scalar terms in the model, each term has a unique grouping factor and the elements of the list returned by `ranef` can be considered as associated with terms or with grouping factors. In more

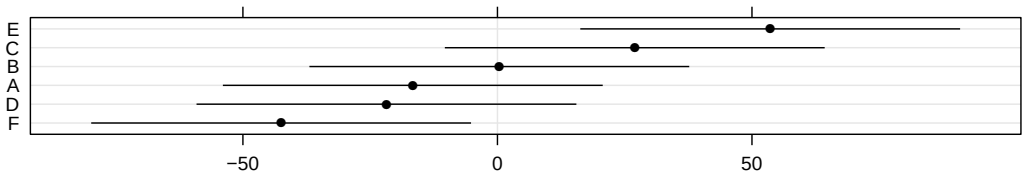
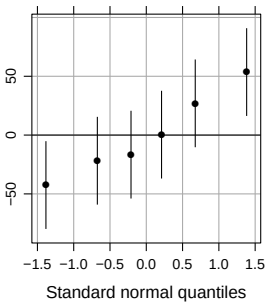


Fig. 1.10 95% prediction intervals on the random effects in `fm1ML`, shown as a dotplot.

Fig. 1.11 95% prediction intervals on the random effects in `fm1ML` versus quantiles of the standard normal distribution.



complex models a particular grouping factor may occur in more than one term, in which case the elements of the list are associated with the grouping factors, not the terms.

Given the data, $\mathbf{y}$, and the parameter estimates, we can evaluate a measure of the dispersion of $(\mathcal{B}|\mathcal{Y} = \mathbf{y})$. In the case of a linear mixed model, this is the conditional standard deviation, from which we can obtain a prediction interval. The `ranef` extractor takes an optional argument, `postVar = TRUE`, which adds these dispersion measures as an attribute of the result. (The name stands for "posterior variance", which is a misnomer that had become established as an argument name before I realized that it wasn't the correct term.)

We can plot these prediction intervals using

```
> dotplot(ranef(fm1ML, postVar = TRUE))
```

(Fig. 1.10), which provides linear spacing of the levels on the y axis, or using

```
> qqmath(ranef(fm1ML, postVar=TRUE))
```

(Fig. 1.11), where the intervals are plotted versus quantiles of the standard normal.

The dotplot is preferred when there are only a few levels of the grouping factor, as in this case. When there are hundreds or thousands of random effects the `qqmath` form is preferred because it focuses attention on the "important few" at the extremes and de-emphasizes the "trivial many" that are close to zero.

1.7 Chapter Summary

A considerable amount of material has been presented in this chapter, especially considering the word "simple" in its title (it's the model that is simple, not the material). A summary may be in order.

A mixed-effects model incorporates fixed-effects parameters and random effects, which are unobserved random variables, $\mathscr{B}$. In a linear mixed model, both the unconditional distribution of $\mathscr{B}$ and the conditional distribution, $(\mathscr{Y}|\mathscr{B}=\mathbf{b})$, are multivariate Gaussian distributions. Furthermore, this conditional distribution is a spherical Gaussian with mean, μ, determined by the linear predictor, $\mathbf{Zb}+\mathbf{X}\beta$. That is,

$$(\mathscr{Y}|\mathscr{B}=\mathbf{b}) \sim \mathscr{N}(\mathbf{Zb}+\mathbf{X}\beta, \sigma^2\mathbf{I}_n).$$

The unconditional distribution of $\mathscr{B}$ has mean $\mathbf{0}$ and a parameterized $q \times q$ variance-covariance matrix, Σ_θ.

In the models we considered in this chapter, Σ_θ, is a simple multiple of the identity matrix, $\mathbf{I}_6$. This matrix is always a multiple of the identity in models with just one random-effects term that is a simple, scalar term. The reason for introducing all the machinery that we did is to allow for more general model specifications.

The maximum likelihood estimates of the parameters are obtained by minimizing the deviance. For linear mixed models we can minimize the profiled deviance, which is a function of θ only, thereby considerably simplifying the optimization problem.

To assess the precision of the parameter estimates, we profile the deviance function with respect to each parameter and apply a signed square root transformation to the likelihood ratio test statistic, producing a profile zeta function for each parameter. These functions provide likelihood-based confidence intervals for the parameters. Profile zeta plots allow us to visually assess the precision of individual parameters. Profile pairs plots allow us to visualize the pairwise dependence of parameter estimates and two-dimensional marginal confidence regions.

Prediction intervals from the conditional distribution of the random effects, given the observed data, allow us to assess the precision of the random effects.

Exercises

These exercises and several others in this book use data sets from the `MEMSS` package for R. You will need to ensure that this package is installed before you can access the data sets.

To load a particular data set, either attach the package

```
> library(MEMSS)
```

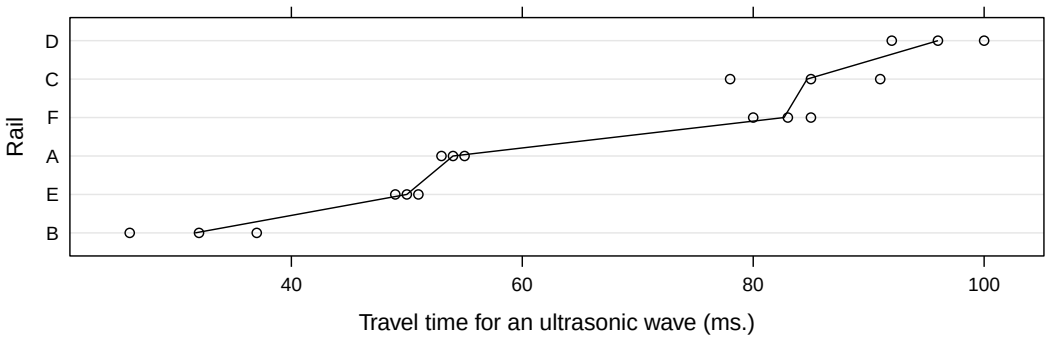


Fig. 1.12 Travel time for an ultrasonic wave test on 6 rails

or load just the one data set

```
> data(Rail, package = "MEMSS")
```

1.1. Check the documentation, the structure (`str`) and a summary of the `Rail` data (Fig. 1.12) from the `MEMSS` package. Note that if you used `data` to access this data set then you must use

```
> help(Rail, package = "MEMSS")
```

to display the documentation for it.

1.2. Fit a model with `travel` as the response and a simple, scalar random-effects term for the variable `Rail`. Use the REML criterion, which is the default. Create a dotplot of the conditional modes of the random effects.

1.3. Refit the model using maximum likelihood. Check the parameter estimates and, in the case of the fixed-effects parameter, its standard error. In what ways have the parameter estimates changed? Which parameter estimates have not changed?

1.4. Profile the fitted model and construct 95% profile-based confidence intervals on the parameters. Is the confidence interval on σ_1 close to being symmetric about the estimate? Is the corresponding interval on $\log(\sigma_1)$ close to being symmetric about its estimate?

1.5. Create the profile zeta plot for this model. For which parameters there good normal approximations?

1.6. Create a profile pairs plot for this model. Does the shape of the deviance contours in this model mirror those in Fig. 1.9?

1.7. Plot the prediction intervals on the random effects from this model. Do any of these prediction intervals contain zero? Consider the relative magnitudes of $\widehat{\sigma_1}$ and $\widehat{\sigma}$ in this model compared to those in model `fm1` for the `Dyestuff` data. Should these ratios of σ_1/σ lead you to expect a different pattern of prediction intervals in this plot than those in Fig. 1.10?

Chapter 2
Models with multiple random-effects terms

The mixed models considered in the previous chapter had only one random-effects term, which was a simple, scalar random-effects term, and a single fixed-effects coefficient. Although such models can be useful, it is with the facility to use multiple random-effects terms and to use random-effects terms beyond a simple, scalar term that we can begin to realize the flexibility and versatility of mixed models.

In this chapter we consider models with multiple simple, scalar random-effects terms, showing examples where the grouping factors for these terms are in completely crossed or nested or partially crossed configurations. For ease of description we will refer to the random effects as being crossed or nested although, strictly speaking, the distinction between nested and non-nested refers to the grouping factors, not the random effects.

2.1 A Model With Crossed Random Effects

One of the areas in which the methods in the `lme4` package for R are particularly effective is in fitting models to cross-classified data where several factors have random effects associated with them. For example, in many experiments in psychology the reaction of each of a set of subjects to each of a group of stimuli or items is measured. If the subjects are considered to be a sample from a population of subjects and the items are a sample from a population of items, then it would make sense to associate random effects with both these factors.

In the past it was difficult to fit mixed models with multiple, crossed grouping factors to large, possibly unbalanced, data sets. The methods in the `lme4` package are able to do this. To introduce the methods let us first consider a small, balanced data set with crossed grouping factors.

2.1.1 The Penicillin Data

The Penicillin data are derived from Table 6.6, p. 144 of Davies and Goldsmith [1972] where they are described as coming from an investigation to

> assess the variability between samples of penicillin by the *B. subtilis* method. In this test method a bulk-innoculated nutrient agar medium is poured into a Petri dish of approximately 90 mm. diameter, known as a plate. When the medium has set, six small hollow cylinders or pots (about 4 mm. in diameter) are cemented onto the surface at equally spaced intervals. A few drops of the penicillin solutions to be compared are placed in the respective cylinders, and the whole plate is placed in an incubator for a given time. Penicillin diffuses from the pots into the agar, and this produces a clear circular zone of inhibition of growth of the organisms, which can be readily measured. The diameter of the zone is related in a known way to the concentration of penicillin in the solution.

As with the Dyestuff data, we examine the structure

```
> str(Penicillin)

'data.frame':        144 obs. of  3 variables:
 $ diameter: num  27 23 26 23 23 21 27 23 26 23 ...
 $ plate   : Factor w/ 24 levels "a","b","c","d",..: 1 1 1 1 1 1 2 2 2 2 ...
 $ sample  : Factor w/ 6 levels "A","B","C","D",..: 1 2 3 4 5 6 1 2 3 4 ...
```

and a summary

```
> summary(Penicillin)

    diameter         plate     sample
 Min.   :18.00   a      :  6   A:24
 1st Qu.:22.00   b      :  6   B:24
 Median :23.00   c      :  6   C:24
 Mean   :22.97   d      :  6   D:24
 3rd Qu.:24.00   e      :  6   E:24
 Max.   :27.00   f      :  6   F:24
                 (Other):108
```

of the Penicillin data, then plot it (Fig. 2.1).

The variation in the diameter is associated with the plates and with the samples. Because each plate is used for only the six samples shown here we will use random effects for the plate. As in the dyestuff example, we are more interested in the sample-to-sample variability in the penicillin samples than in the potency of a particular sample. Hence we will also use random effects for the sample.

In this experiment each sample is used on each plate. We say that the sample and plate factors are *crossed*, as opposed to *nested* factors, which we will describe in the next section. By itself, the designation "crossed" just means that the factors are not nested. If we wish to be more specific, we could describe these factors as being *completely crossed*, which means that we have at least one observation for each combination of a level of sample and

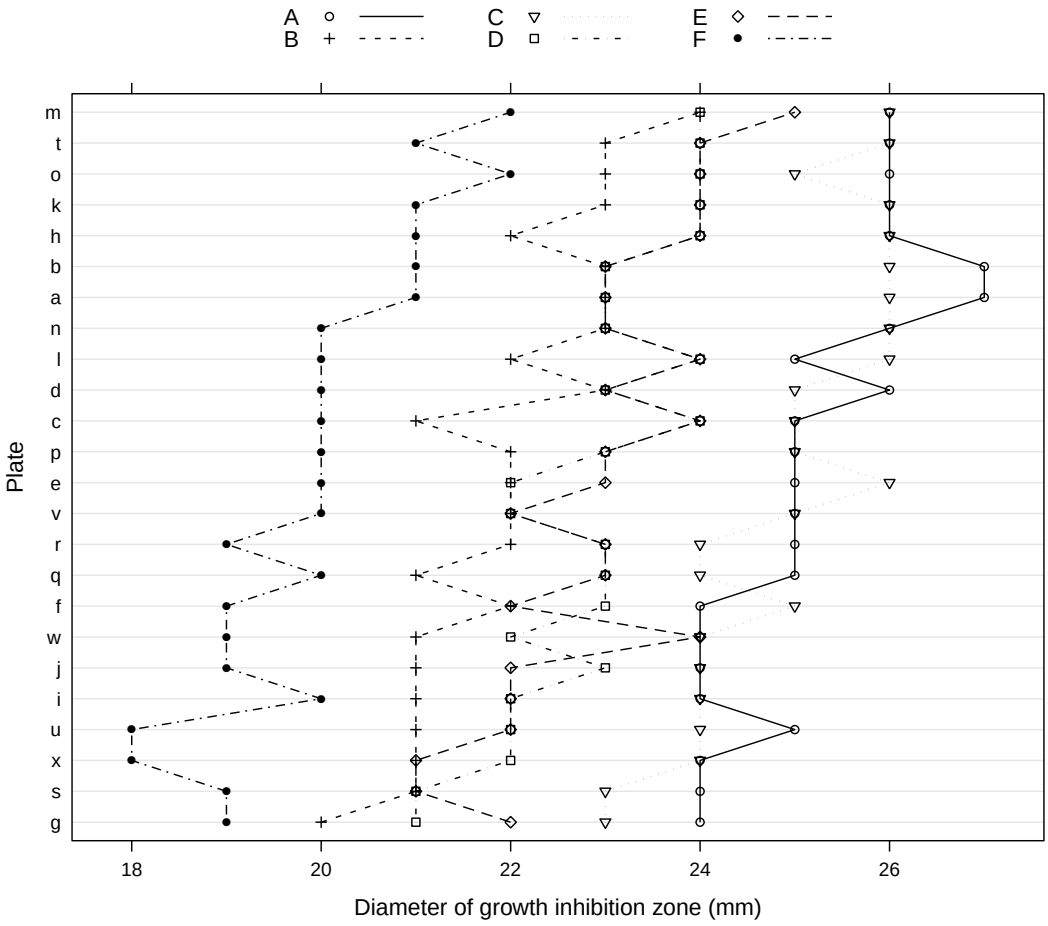


Fig. 2.1 Diameter of the growth inhibition zone (mm) in the *B. subtilis* method of assessing the concentration of penicillin. Each of 6 samples was applied to each of the 24 agar plates. The lines join observations on the same sample.

a level of `plate`. We can see this in Fig. 2.1 and, because there are moderate numbers of levels in these factors, we can check it in a cross-tabulation

```
> xtabs(~ sample + plate, Penicillin)
```

```
      plate
sample a b c d e f g h i j k l m n o p q r s t u v w x
     A 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
     B 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
     C 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
     D 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
     E 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
     F 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

Like the `Dyestuff` data, the factors in the `Penicillin` data are balanced. That is, there are exactly the same number of observations on each plate and for each sample and, furthermore, there is the same number of observations on each combination of levels. In this case there is exactly one observation for

each combination of sample and plate. We would describe the configuration of these two factors as an unreplicated, completely balanced, crossed design.

In general, balance is a desirable but precarious property of a data set. We may be able to impose balance in a designed experiment but we typically cannot expect that data from an observation study will be balanced. Also, as anyone who analyzes real data soon finds out, expecting that balance in the design of an experiment will produce a balanced data set is contrary to "Murphy's Law". That's why statisticians allow for missing data. Even when we apply each of the six samples to each of the 24 plates, something could go wrong for one of the samples on one of the plates, leaving us without a measurement for that combination of levels and thus an unbalanced data set.

2.1.2 A Model for the `Penicillin` *Data*

A model incorporating random effects for both the `plate` and the `sample` is straightforward to specify — we include simple, scalar random effects terms for both these factors.

```
> (fm2 <- lmer(diameter ~ 1 + (1|plate) + (1|sample), Penicillin))
```

```
Linear mixed model fit by REML
Formula: diameter ~ 1 + (1 | plate) + (1 | sample)
   Data: Penicillin
 REML
330.9

Random effects:
 Groups   Name        Variance Std.Dev.
 plate    (Intercept) 0.71691  0.84671
 sample   (Intercept) 3.73097  1.93157
 Residual             0.30241  0.54992
Number of obs: 144, groups: plate, 24; sample, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)  22.9722     0.8086   28.41
```

This model display indicates that the sample-to-sample variability has the greatest contribution, then plate-to-plate variability and finally the "residual" variability that cannot be attributed to either the sample or the plate. These conclusions are consistent with what we see in the `Penicillin` data plot (Fig. 2.1).

The prediction intervals on the random effects (Fig. 2.2) confirm that the conditional distribution of the random effects for `sample` has much less variability than does the conditional distribution of the random effects for `sample`.

In chapter 1 we saw that a model with a single, simple, scalar random-effects term generated a random-effects model matrix, $\mathbf{Z}$, that is the matrix of

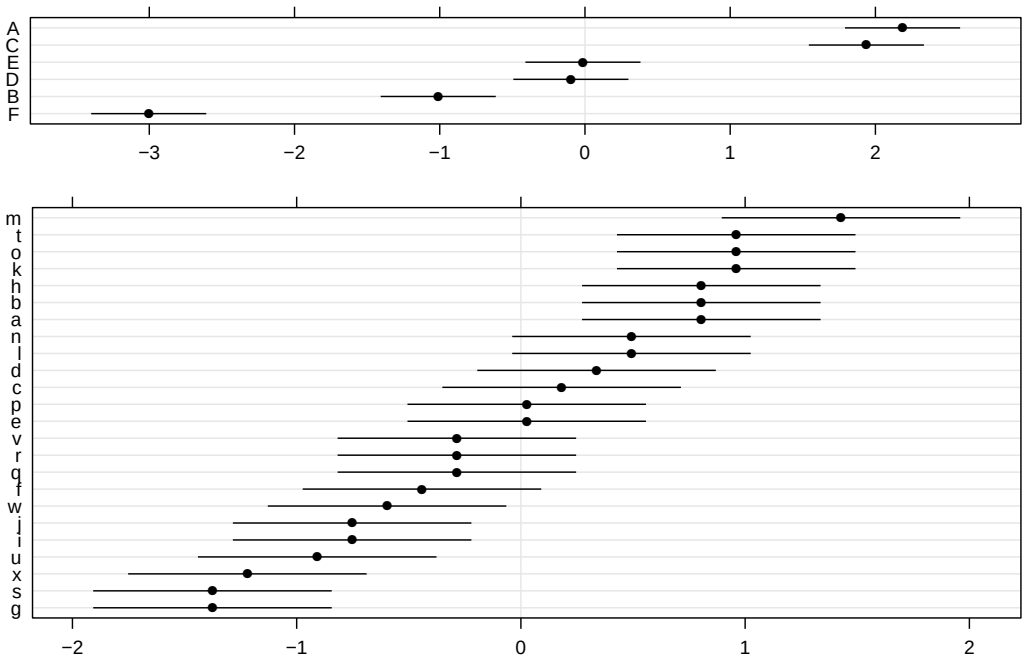


Fig. 2.2 95% prediction intervals on the random effects for model `fm2` fit to the `Penicillin` data.

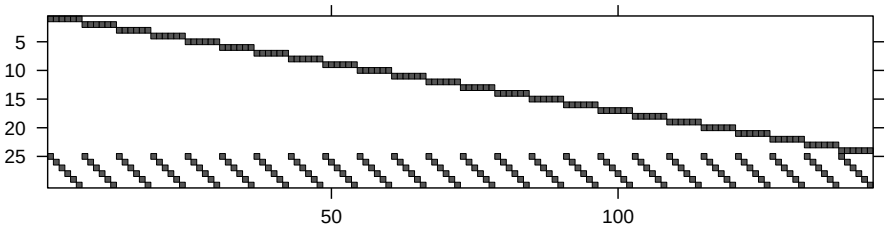


Fig. 2.3 Image of the transpose of the random-effects model matrix, **Z**, for model `fm2`. The non-zero elements, which are all unity, are shown as darkened squares. The zero elements are blank.

indicators of the levels of the grouping factor. When we have multiple, simple, scalar random-effects terms, as in model `fm2`, each term generates a matrix of indicator columns and all the sets of indicator columns are concatenated to form the model matrix **Z**. The transpose of this matrix contains rows of indicators for each factor, as shown in Fig. 2.3.

The relative covariance factor (Fig. 2.4, left panel) is no longer a multiple of the identity. It is now block diagonal, with two blocks, one of size 24 and one of size 6, each of which is a multiple of the identity. The diagonal elements in each block are θ_1 and θ_2. The numeric values of these parameters can be obtained as

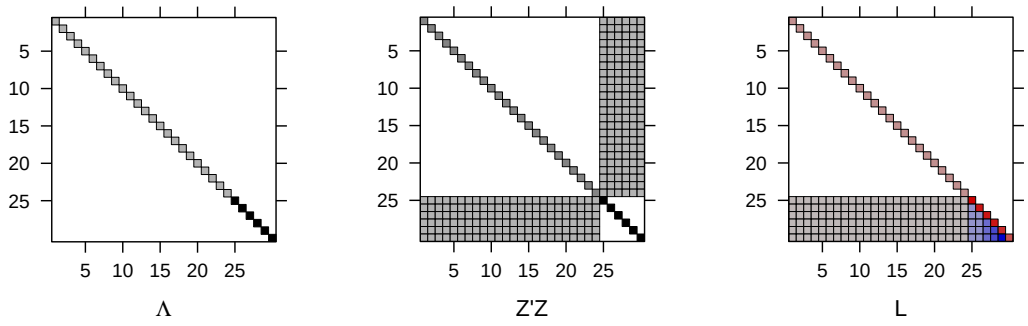


Fig. 2.4 Images of the relative covariance factor, Λ, the cross-product of the random-effects model matrix, $\mathbf{Z}^\mathsf{T}\mathbf{Z}$, and the sparse Cholesky factor, $\mathbf{L}$, for model `fm2`.

```
> env(fm2)$theta

[1] 1.539683 3.512443
```

The first parameter is the relative standard deviation of the random effects for `plate`, which has the value (0.84671/0.54992) at convergence, and the second is the relative standard deviation of the random effects for `sample` (1.93157/0.54992).

Because Λ_θ is diagonal, the pattern of non-zeros in $\Lambda_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{Z}\Lambda_\theta + \mathbf{I}$ will be the same as that in $\mathbf{Z}^\mathsf{T}\mathbf{Z}$, shown in the middle panel of Fig. 2.4. The sparse Cholesky factor, $\mathbf{L}$, shown in the right panel is lower triangular and has non-zero elements in the lower right hand corner in positions where $\mathbf{Z}^\mathsf{T}\mathbf{Z}$ has systematic zeros. We say that "fill-in" has occurred when forming the sparse Cholesky decomposition. In this case there is a relatively minor amount of fill but in other cases there can be a substantial amount of fill and we shall take precautions so as to reduce this, because fill-in adds to the computational effort in determining the MLEs or the REML estimates.

A profile zeta plot (Fig. 2.5) for the parameters in model `fm2` leads to conclusions similar to those from Fig. 1.5 for model `fm1ML` in the previous chapter. The fixed-effect parameter, β_0, for the `(Intercept)` term has symmetric intervals and is over-dispersed relative to the normal distribution. The logarithm of σ has a good normal approximation but the standard deviations of the random effects, σ_1 and σ_2, are skewed. The skewness for σ_2 is worse than that for σ_1, because the estimate of σ_2 is less precise than that of σ_1, in both absolute and relative senses. For an absolute comparison we compare the widths of the confidence intervals for these parameters.

```
> confint(pr2)

                2.5 %     97.5 %
.sig01      0.6335658  1.1821040
```

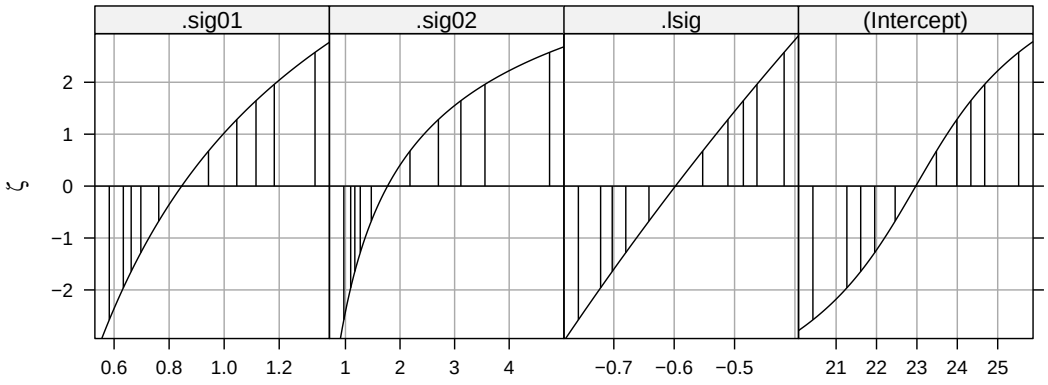


Fig. 2.5 Profile zeta plot of the parameters in model `fm2`.

```
.sig02       1.0957822  3.5563194
.lsig       -0.7218645 -0.4629033
(Intercept) 21.2666274 24.6778176
```

In a relative comparison we examine the ratio of the endpoints of the interval divided by the estimate.

```
> confint(pr2)[1:2,]/c(0.8455722, 1.770648)

           2.5 %   97.5 %
.sig01 0.7492746 1.397993
.sig02 0.6188594 2.008485
```

The lack of precision in the estimate of σ_2 is a consequence of only having 6 distinct levels of the `sample` factor. The `plate` factor, on the other hand, has 24 distinct levels. In general it is more difficult to estimate a measure of spread, such as the standard deviation, than to estimate a measure of location, such as a mean, especially when the number of levels of the factor is small. Six levels are about the minimum number required for obtaining sensible estimates of standard deviations for simple, scalar random effects terms.

The profile pairs plot (Fig. 2.6) shows patterns similar to those in Fig. 1.9 for pairs of parameters in model `fm1` fit to the `Dyestuff` data. On the ζ scale (panels below the diagonal) the profile traces are nearly straight and orthogonal with the exception of the trace of $\zeta(\sigma_2)$ on $\zeta(\beta_0)$ (the horizontal trace for the panel in the $(4,2)$ position. The pattern of this trace is similar to the pattern of the trace of $\zeta(\sigma_1)$ on $\zeta(\beta_0)$ in Fig. 1.9. Moving β_0 from its estimate, $\widehat{\beta}_0$ in either direction will increase the residual sum of squares. The increase in the residual variability is reflected in an increase of one or more of the dispersion parameters. The balanced experimental design results in a fixed estimate of σ and the extra apparent variability must be incorporated into σ_1 or σ_2.

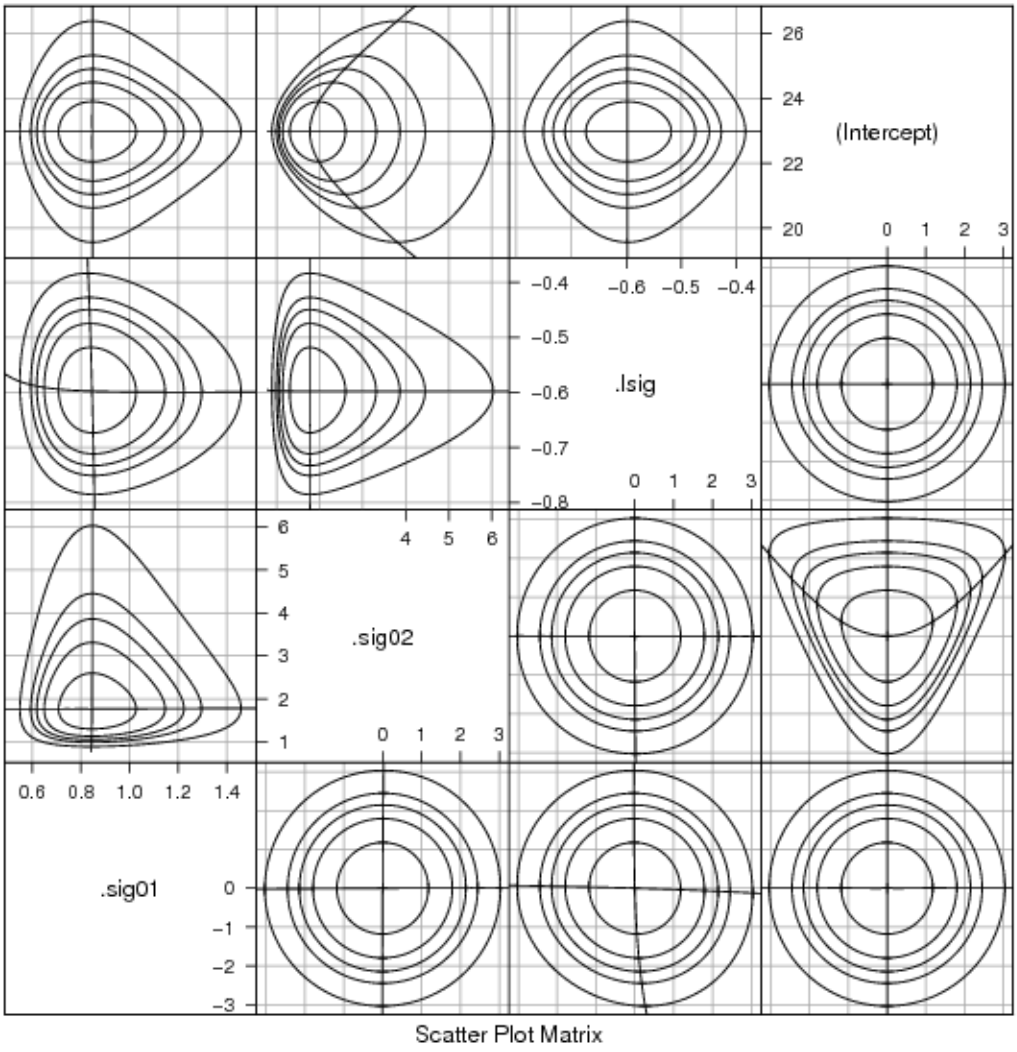


Fig. 2.6 Profile pairs plot for the parameters in model `fm2` fit to the `Penicillin` data.

Contours in panels of parameter pairs on the original scales (i.e. panels above the diagonal) can show considerable distortion from the ideal elliptical shape. For example, contours in the σ_2 versus σ_1 panel (the $(1,2)$ position) and the $\log(\sigma)$ versus σ_2 panel (in the $(2,3)$ position) are dramatically non-elliptical. However, the distortion of the contours is not due to these parameter estimates depending on each other. It is almost entirely due to the choice of scale for σ_1 and σ_2. When we plot the contours on the scale of $\log(\sigma_1)$ and $\log(\sigma_2)$ instead (Fig. **??**) they are much closer to the elliptical pattern.

Conversely, if we tried to plot contours on the scale of σ_1^2 and σ_2^2 (not shown), they would be hideously distorted.

2.2 A Model With Nested Random Effects

In this section we again consider a simple example, this time fitting a model with *nested* grouping factors for the random effects.

2.2.1 The `Pastes` *Data*

The third example from Davies and Goldsmith [1972, Table 6.5, p. 138] is described as coming from

> deliveries of a chemical paste product contained in casks where, in addition to sampling and testing errors, there are variations in quality between deliveries ... As a routine, three casks selected at random from each delivery were sampled and the samples were kept for reference. ... Ten of the delivery batches were sampled at random and two analytical tests carried out on each of the 30 samples.

The structure and summary of the `Pastes` data object are

```
> str(Pastes)

'data.frame':        60 obs. of  4 variables:
 $ strength: num  62.8 62.6 60.1 62.3 62.7 63.1 60 61.4 57.5 56.9 ...
 $ batch   : Factor w/ 10 levels "A","B","C","D",..: 1 1 1 1 1 1 2 2 2 2 ...
 $ cask    : Factor w/ 3 levels "a","b","c": 1 1 2 2 3 3 1 1 2 2 ...
 $ sample  : Factor w/ 30 levels "A:a","A:b","A:c",..: 1 1 2 2 3 3 4 4 5 5 ...
```

```
> summary(Pastes)

    strength         batch    cask       sample
 Min.   :54.20   A      : 6   a:20   A:a    : 2
 1st Qu.:57.50   B      : 6   b:20   A:b    : 2
 Median :59.30   C      : 6   c:20   A:c    : 2
 Mean   :60.05   D      : 6          B:a    : 2
 3rd Qu.:62.88   E      : 6          B:b    : 2
 Max.   :66.00   F      : 6          B:c    : 2
                 (Other):24          (Other):48
```

As stated in the description in Davies and Goldsmith [1972], there are 30 samples, three from each of the 10 delivery batches. We have labelled the levels of the `sample` factor with the label of the `batch` factor followed by 'a', 'b' or 'c' to distinguish the three samples taken from that batch. The cross-tabulation produced by the `xtabs` function, using the optional argument `sparse = TRUE`, provides a concise display of the relationship.

```
> xtabs(~ batch + sample, Pastes, drop = TRUE, sparse = TRUE)

10 x 30 sparse Matrix of class "dgCMatrix"

A 2 2 2 . . . . . .
B . . . 2 2 2 . . . .
```

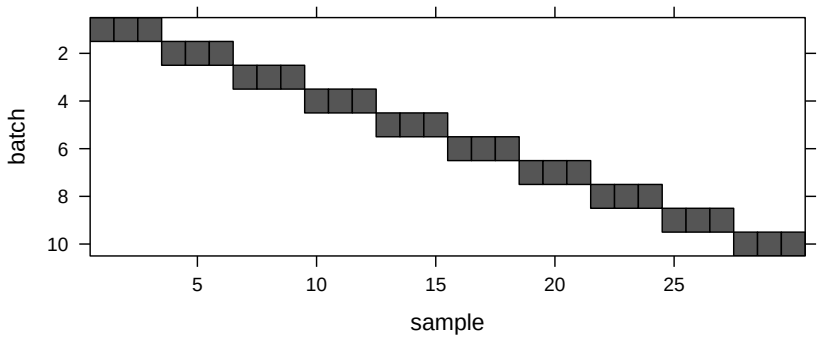


Fig. 2.7 Image of the cross-tabulation of the `batch` and `sample` factors in the `Pastes` data.

```
C . . . . . 2 2 2 . . . . . .
D . . . . . 2 2 2 . . . . . .
E . . . . . 2 2 2 . . . . . .
F . . . . . 2 2 2 . . . . . .
G . . . . . 2 2 2 . . . . . .
H . . . . . 2 2 2 . . . . . .
I . . . . . 2 2 2 . . . . . .
J . . . . . 2 2 2
```

Alternatively, we can use an image (Fig. 2.7) of this cross-tabulation to visualize the structure.

When plotting the `strength` versus `batch` and `sample` in the `Pastes` data we should remember that we have two strength measurements on each of the 30 samples. It is tempting to use the cask designation ('a', 'b' and 'c') to determine, say, the plotting symbol within a `batch`. It would be fine to do this within a batch but the plot would be misleading if we used the same symbol for cask 'a' in different batches. There is no relationship between cask 'a' in batch 'A' and cask 'a' in batch 'B'. The labels 'a', 'b' and 'c' are used only to distinguish the three samples within a batch; they do not have a meaning across batches.

In Fig. 2.8 we plot the two strength measurements on each of the samples within each of the batches and join up the average strength for each sample. The perceptive reader will have noticed that the levels of the factors on the vertical axis in this figure, and in Fig. 1.1 and 2.1, have been reordered according to increasing average response. In all these cases there is no inherent ordering of the levels of the covariate such as `batch` or `plate`. Rather than confuse our interpretation of the plot by determining the vertical displacement of points according to a random ordering, we impose an ordering according to increasing mean response. This allows us to more easily check for structure in the data, including undesirable characteristics like increasing variability of the response with increasing mean level of the response.

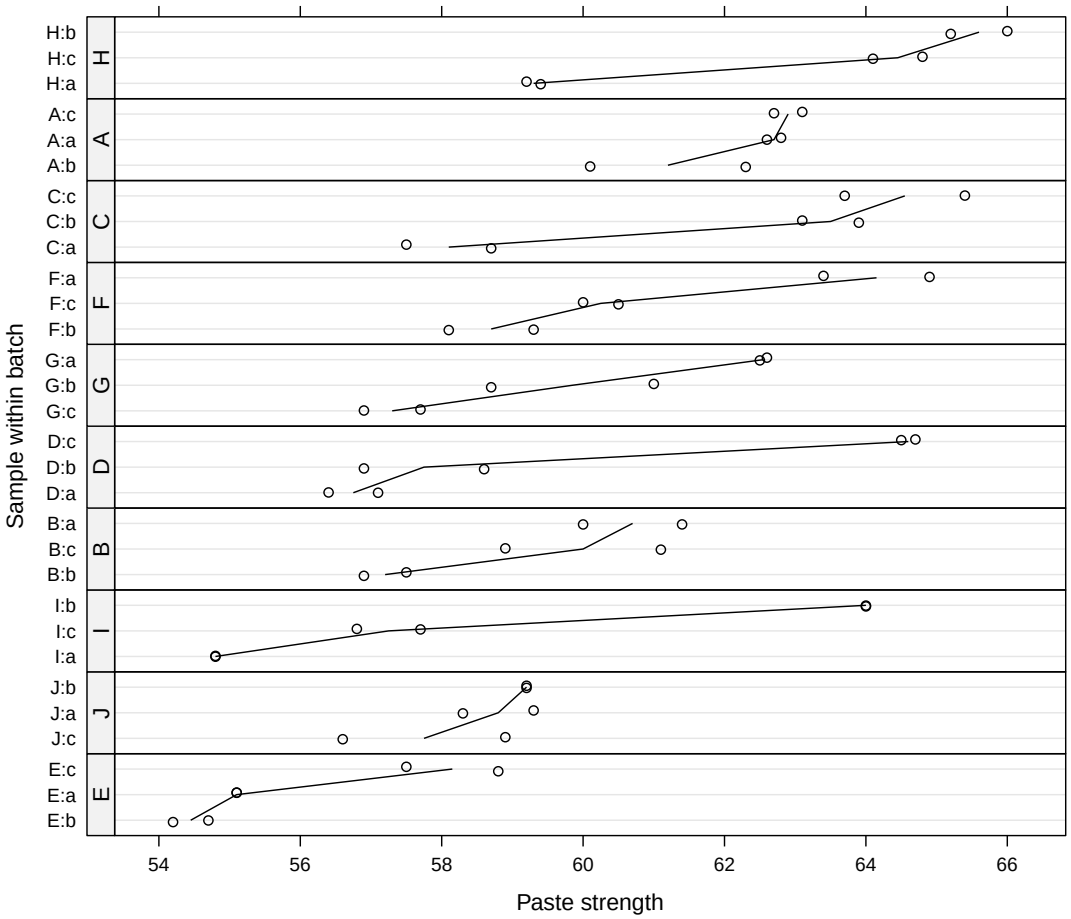


Fig. 2.8 Strength of paste preparations according to the `batch` and the `sample` within the batch. There were two strength measurements on each of the 30 samples; three samples each from 10 batches.

In Fig. 2.8 we order the samples within each batch separately then order the batches according to increasing mean strength.

Figure 2.8 shows considerable variability in strength between samples relative to the variability within samples. There is some indication of variability between batches, in addition to the variability induced by the samples, but not a strong indication of a batch effect. For example, each of batches I and D, with low mean strength relative to the other batches, contained one sample (I:b and D:c, respectively) that had high mean strength relative to the other samples. Also, batches H and C, with comparatively high mean batch strength. contain samples H:a and C:a with comparatively low mean sample strength. In Sect. 2.2.4 we will examine the need for incorporating batch-to-batch variability in addition to sample-to-sample variability in the statistical model.

2.2.1.1 Nested Factors

Because each level of `sample` occurs with one and only one level of `batch` we say that `sample` is *nested within* `batch`. Some presentations of mixed-effects models, especially those related to *multilevel modeling* [Rasbash et al., 2000] or *hierarchical linear models* [Raudenbush and Bryk, 2002], leave the impression that one can only define random effects with respect to factors that are nested. This is the origin of the terms "multilevel", referring to multiple, nested levels of variability, and "hierarchical", also invoking the concept of a hierarchy of levels. To be fair, both those references do describe the use of models with random effects associated with non-nested factors, but such models tend to be treated as a special case.

The blurring of mixed-effects models with the concept of multiple, hierarchical levels of variation results in an unwarranted emphasis on "levels" when defining a model and leads to considerable confusion. It is perfectly legitimate to define models having random effects associated with non-nested factors. The reasons for the emphasis on defining random effects with respect to nested factors only are that such cases do occur frequently in practice and that some of the computational methods for estimating the parameters in the models can only be easily applied to nested factors.

This is not the case for the methods used in the `lme4` package. Indeed there is nothing special done for models with random effects for nested factors. When random effects are associated with multiple factors exactly the same computational methods are used whether the factors form a nested sequence or are partially crossed or are completely crossed. A case of a nested sequence of "grouping factors" for the random effects (including the trivial case of only one such factor) is detected but this information does not change the course of the computation. It is available to be used as a diagnostic check. When the user knows that the grouping factors should be nested, she can check if they are indeed nested.

There is, however, one aspect of nested grouping factors that we should emphasize, which is the possibility of a factor that is *implicitly nested* within another factor. Suppose, for example, that the `sample` factor was defined as having three levels instead of 30 with the implicit assumption that `sample` is nested within `batch`. It may seem silly to try to distinguish 30 different batches with only three levels of a factor but, unfortunately, data are frequently organized and presented like this, especially in text books. The `cask` factor in the `Pastes` data is exactly such an implicitly nested factor. If we cross-tabulate `batch` and `cask`

```
> xtabs(~ cask + batch, Pastes)
```

```
    batch
cask A B C D E F G H I J
   a 2 2 2 2 2 2 2 2 2 2
   b 2 2 2 2 2 2 2 2 2 2
   c 2 2 2 2 2 2 2 2 2 2
```

are crossed, not nested. If we know that the cask should be considered as nested within the batch then we should create a new categorical variable giving the batch-cask combination, which is exactly what the `sample` factor is. A simple way to create such a factor is to use the interaction operator, ':', on the factors. It is advisable, but not necessary, to apply `factor` to the result thereby dropping unused levels of the interaction from the set of all possible levels of the factor. (An "unused level" is a combination that does not occur in the data.) A convenient code idiom is

```
> Pastes$sample <- with(Pastes, factor(batch:cask))
```

or

```
> Pastes <- within(Pastes, sample <- factor(batch:cask))
```

In a small data set like `Pastes` we can quickly detect a factor being implicitly nested within another factor and take appropriate action. In a large data set, perhaps hundreds of thousands of test scores for students in thousands of schools from hundreds of school districts, it is not always obvious if school identifiers are unique across the entire data set or just within a district. If you are not sure, the safest thing to do is to create the interaction factor, as shown above, so you can be confident that levels of the district:school interaction do indeed correspond to unique schools.

2.2.2 Fitting a Model With Nested Random Effects

Fitting a model with simple, scalar random effects for nested factors is done in exactly the same way as fitting a model with random effects for crossed grouping factors. We include random-effects terms for each factor, as in

```
> (fm3 <- lmer(strength ~ 1 + (1|sample) + (1|batch), Pastes, REML=0))

Linear mixed model fit by maximum likelihood
Formula: strength ~ 1 + (1 | sample) + (1 | batch)
   Data: Pastes
 AIC   BIC logLik deviance
 256 264.4   -124      248

Random effects:
 Groups   Name        Variance Std.Dev.
 sample   (Intercept) 8.4337   2.9041
 batch    (Intercept) 1.1992   1.0951
 Residual             0.6780   0.8234
Number of obs: 60, groups: sample, 30; batch, 10

Fixed effects:
            Estimate Std. Error t value
(Intercept)  60.0533     0.6421   93.52
```

Not only is the model specification similar for nested and crossed factors,

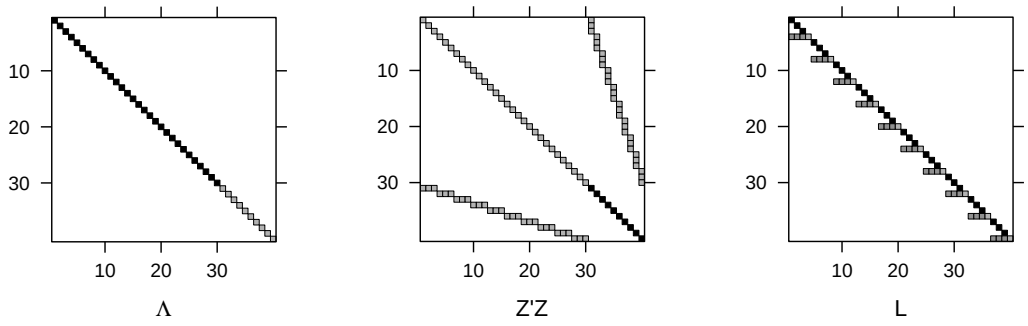


Fig. 2.9 Images of the relative covariance factor, Λ, the cross-product of the random-effects model matrix, $\mathbf{Z}^\mathsf{T}\mathbf{Z}$, and the sparse Cholesky factor, $\mathbf{L}$, for model `fm3`.

the internal calculations are performed according to the methods described in Sect. 1.4.1 for each model type. Comparing the patterns in the matrices Λ, $\mathbf{Z}^\mathsf{T}\mathbf{Z}$ and $\mathbf{L}$ for this model (Fig. 2.9) to those in Fig. 2.4 shows that models with nested factors produce simple repeated structures along the diagonal of the sparse Cholesky factor, $\mathbf{L}$, after reordering the random effects (we discuss this reordering later in Sect. **??**). This type of structure has the desirable property that there is no "fill-in" during calculation of the Cholesky factor. In other words, the number of non-zeros in $\mathbf{L}$ is the same as the number of non-zeros in the lower triangle of the matrix being factored, $\Lambda^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{Z}\Lambda + \mathbf{I}$ (which, because Λ is diagonal, has the same structure as $\mathbf{Z}^\mathsf{T}\mathbf{Z}$).

Fill-in of the Cholesky factor is not an important issue when we have a few dozen random effects, as we do here. It is an important issue when we have millions of random effects in complex configurations, as has been the case in some of the models that have been fit using `lmer`.

2.2.3 Assessing the Parameter Estimates for Model `fm3`

The parameter estimates are: $\widehat{\sigma_1}$ =2.904, the standard deviation of the random effects for `sample`; $\widehat{\sigma_2}$ =1.095, the standard deviation of the random effects for `batch`; $\widehat{\sigma}$ =0.823, the standard deviation of the residual noise term; and $\widehat{\beta_0}$ =60.053, the overall mean response, which is labeled `(Intercept)` in these models.

The estimated standard deviation for `sample` is nearly three times as large as that for `batch`, which confirms what we saw in Fig. 2.8. Indeed our conclusion from Fig. 2.8 was that there may not be a significant batch-to-batch variability in addition to the sample-to-sample variability.

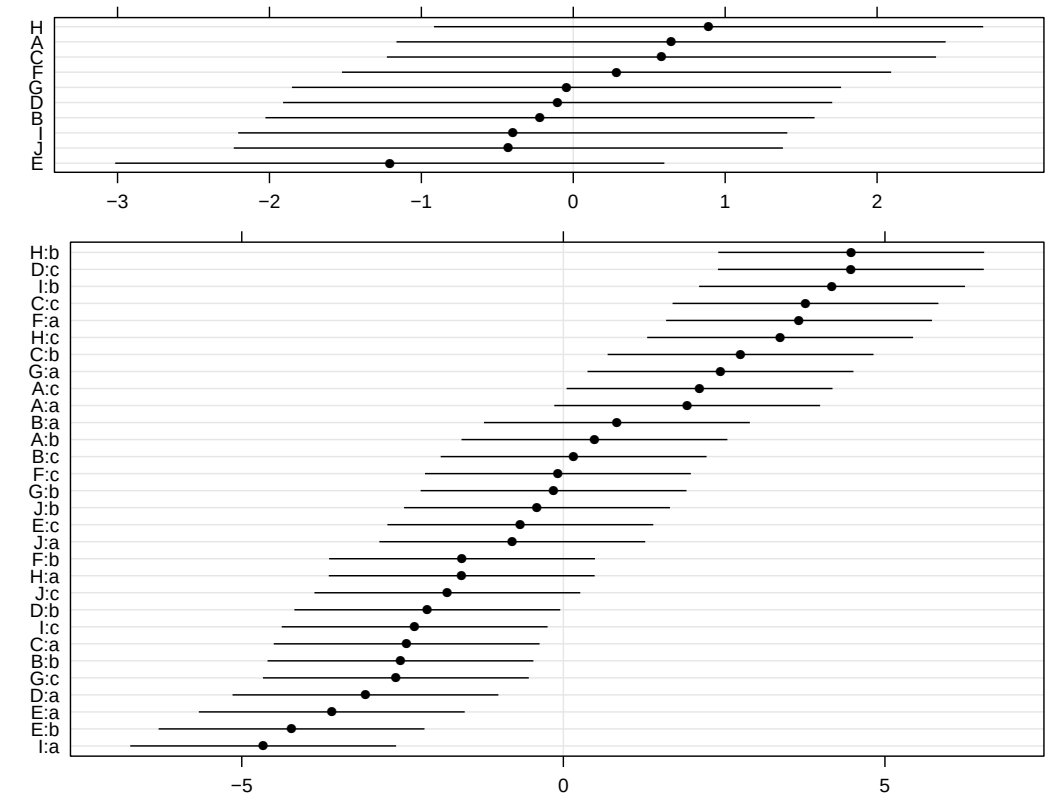


Fig. 2.10 95% prediction intervals on the random effects for model `fm2` fit to the `Penicillin` data.

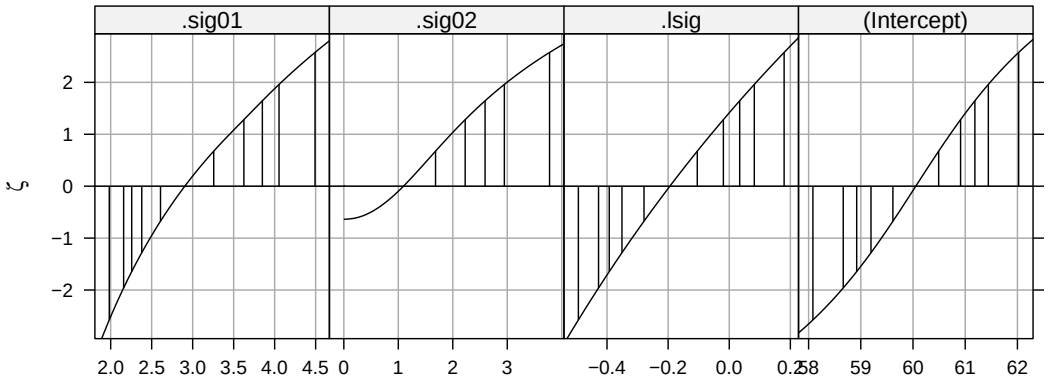


Fig. 2.11 Profile zeta plots for the parameters in model `fm3`.

Plots of the prediction intervals of the random effects (Fig. 2.10)) confirm this impression in that all the prediction intervals for the random effects for `batch` contain zero. Furthermore, the profile zeta plot (Fig. 2.11) shows that the even the 50% profile-based confidence interval on σ_2 extends to zero.

2.2.4 Testing $H_0 : \sigma_2 = 0$ Versus $H_a : \sigma_2 > 0$

One of the many famous quotes from Albert Einstein is "Everything should be made as simple as possible, but not simpler." In statistical modeling this *principal of parsimony* is embodied in hypothesis test comparing two models, one of which contains the other as a special case. Typically, one or more of the parameters in the more general model, which we call the *alternative hypothesis*, is constrained in some way, resulting in the restricted model, which we call the *null hypothesis*. Although we phrase the hypothesis test in terms of the parameter restriction, it is important to realize that we are comparing the quality of fits obtained with two different models.

Because the more general model, H_a, must provide a fit that is at least as good as the restricted model, H_0, our purpose is to determine whether the change in the quality of the fit is sufficient to justify the greater complexity of model H_0. This comparison is often reduced to a *p-value* which is the probability of seeing a difference in the model fits as large as we did, or even larger, when, in fact, H_0 is adequate. Like all probabilities, a p-value must be between 0 and 1. When the p-value for a test is small (close to zero) we prefer the more complex model, saying that we "reject H_0 in favor of H_a". On the other hand, when the p-value is not small we "fail to reject H_0", arguing that there is a non-negligible probability that the observed difference in the model fits could reasonably be the result of random chance, not the inherent superiority of the model H_a. Under these circumstances we prefer the simpler model, H_0 according to the principal of parsimony.

These are the general principles of statistical hypothesis tests. To perform a test in practice we must specify the criterion for comparing the model fits, the method for calculating the p-value from the observed value of the criterion and the standard by which we will determine if the p-value is "small" or not. The criterion is called the *test statistic*, the p-value is calculated from a *reference distribution* for the test statistic, and the standard for small p-values is called the *level* of the test.

In Sect. 1.5 we referred to likelihood ratio tests (LRTs) for which the test statistic is the difference in the deviance. That is, the LRT statistic is $d_0 - d_a$ where d_a is the deviance in the more general (H_a) model fit and d_0 is the deviance in the constrained (H_0) model. An approximate reference distribution for an LRT statistic is the χ^2_ν distribution where ν, the degrees of freedom, is determined by the number of constraints imposed on the parameters of H_a to produce H_0.

The restricted model fit

```
> (fm3a <- lmer(strength ~ 1 + 1|sample, Pastes, REML=0))

Linear mixed model fit by maximum likelihood
Formula: strength ~ 1 + 1 | sample
   Data: Pastes
   AIC   BIC logLik deviance
 254.4 260.7 -124.2    248.4
```

```
Random effects:
 Groups   Name        Variance Std.Dev.
 sample   (Intercept) 9.6328   3.1037
 Residual             0.6780   0.8234
Number of obs: 60, groups: sample, 30

Fixed effects:
            Estimate Std. Error t value
(Intercept)  60.0533     0.5765   104.2
```

is compared to model `fm3` with the `anova` function

```
> anova(fm3a, fm3)

Data: Pastes
Models:
fm3a: strength ~ 1 + 1 | sample
fm3: strength ~ 1 + (1 | sample) + (1 | batch)
     Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
fm3a  3 254.40 260.69 -124.20
fm3   4 255.99 264.37 -124.00 0.4072      1     0.5234
```

which provides a p-value of 0.5234.Because typical standards for "small" p-values are 5% or 1%, a p-value over 50% would not be considered significant at any reasonable level.

We do need to be cautious in quoting this p-value, however, because the parameter value being tested, $\sigma_2 = 0$ is on the boundary of set of possible values, $\sigma_2 \geq 0$, for this parameter. The argument for using a χ^2_1 distribution to calculate a p-value for the change in the deviance does not apply when the parameter value being tested is on the boundary. As shown in Pinheiro and Bates [2000, Sect. 2.5], the p-value from the χ^2_1 distribution will be "conservative" in the sense that it is larger than a simulation-based p-value would be. In the worst-case scenario the χ^2-based p-value will be twice as large as it should be but, even if that were true, an effective p-value of 26% would not cause us to reject H_0 in favor of H_a.

2.2.5 Assessing the Reduced Model, `fm3a`

The profile zeta plots of the parameters in model `fm3a` (Fig. 2.12) are similar to the corresponding plots in Fig. 2.11, as confirmed by the numerical values of the confidence intervals.

```
> confint(pr3)

                 2.5 %      97.5 %
.sig01       2.1579337  4.05358895
.sig02              NA  2.94658928
.lsig       -0.4276761  0.08199287
(Intercept) 58.6636504 61.44301637
```

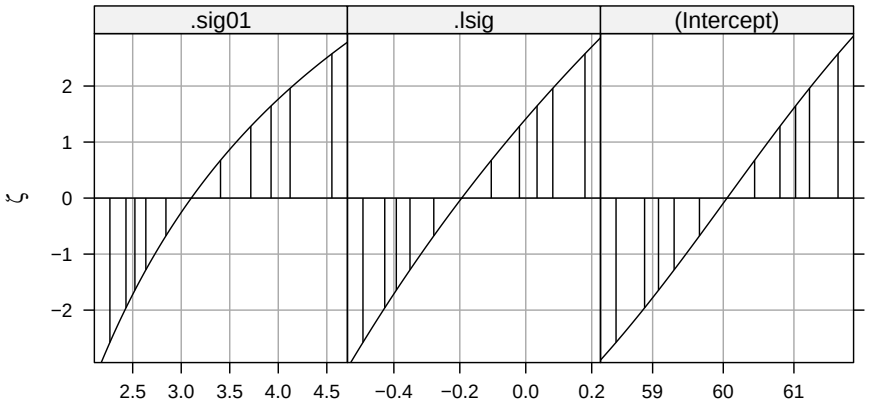


Fig. 2.12 Profile zeta plots for the parameters in model `fm3a`.

```
> confint(pr3a)

                  2.5 %      97.5 %
.sig01        2.4306377  4.12201052
.lsig        -0.4276772  0.08199277
(Intercept) 58.8861831 61.22048353
```

The confidence intervals on $\log(\sigma)$ and β_0 are similar for the two models. The confidence interval on σ_1 is slightly wider in model `fm3a` than in `fm3`, because the variability that is attributed to `batch` in `fm3` is incorporated into the variability due to `sample` in `fm3a`.

The patterns in the profile pairs plot (Fig. 2.13) for the reduced model `fm3a` are similar to those in Fig. 1.9, the profile pairs plot for model `fm1`.

2.3 A Model With Partially Crossed Random Effects

Especially in observational studies with multiple grouping factors, the configuration of the factors is neither nested nor completely crossed. We describe such situations as having *partially crossed* grouping factors for the random effects.

Studies in education, in which test scores for students over time are also associated with teachers and schools, usually result in partially crossed grouping factors. When students with scores in different years have different teachers for the different years, the student factor is not nested within the teacher factor. To have complete crossing of the student and teacher factor is would be necessary for each student to be observed with each teacher which would be unusual. A study of thousands of students and hundreds of teachers inevitably ends up partially crossed.

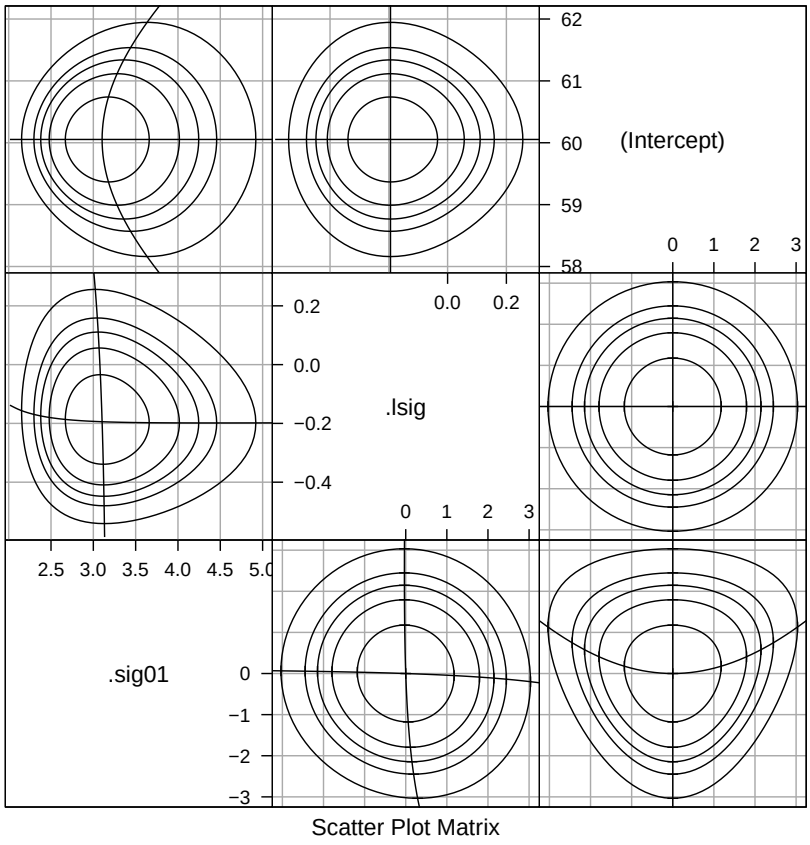


Fig. 2.13 Profile pairs plot for the parameters in model `fm3a` fit to the `Pastes` data.

In this section we consider an example with thousands of students and instructors where the response is the student's evaluation of the instructor's effectiveness. These data, like those from most large observational studies, are quite unbalanced.

2.3.1 The `InstEval` *Data*

The `InstEval` data are from a special evaluation of lecturers by students at ETH-Zürich, to determine who should receive the "best-liked professor" award. These data have been slightly simplified and identifying labels removed so as to preserve anonymity.

The variables

```
> str(InstEval)

'data.frame':        73421 obs. of  7 variables:
 $ s      : Factor w/ 2972 levels "1","2","3","4",..: 1 1 1 1 2 2 3 3 3 3 ...
 $ d      : Factor w/ 1128 levels "1","6","7","8",..: 525 560 832 1068 62 406 3 6 19 75 ...
 $ studage: Ord.factor w/ 4 levels "2"<"4"<"6"<"8": 1 1 1 1 1 1 1 1 1 1 ...
```

```
 $ lectage: Ord.factor w/ 6 levels "1"<"2"<"3"<"4"<..: 2 1 2 2 1 1 1 1 1 1 ...
 $ service: Factor w/ 2 levels "0","1": 1 2 1 2 1 1 2 1 1 1 ...
 $ dept   : Factor w/ 14 levels "15","5","10",..: 14 5 14 12 2 2 13 3 3 3 ...
 $ y      : int  5 2 5 3 2 4 4 5 5 4 ...
```

have somewhat cryptic names. Factor `s` designates the student and `d` the instructor. The `dept` factor is the department for the course and `service` indicates whether the course was a service course taught to students from other departments.

Although the response, `y`, is on a scale of 1 to 5,

```
> xtabs(~ y, InstEval)

y
    1     2     3     4     5
10186 12951 17609 16921 15754
```

it is sufficiently spread out to warrant treating is as if it were a continuous response.

In this chapter we fit models that have random effects for student, instructor, and department or the `dept:service` combination to these data. Later we will fit models that have fixed for instructor and department to these data.

```
> (fm4 <- lmer(y ~ 1 + (1|s) + (1|d) + (1|dept:service), InstEval, REML=0))

Linear mixed model fit by maximum likelihood
Formula: y ~ 1 + (1 | s) + (1 | d) + (1 | dept:service)
   Data: InstEval
    AIC    BIC  logLik deviance
 237663 237709 -118827   237653

Random effects:
 Groups       Name        Variance Std.Dev.
 s            (Intercept) 0.105404 0.32466
 d            (Intercept) 0.262563 0.51241
 dept:service (Intercept) 0.012126 0.11012
 Residual                 1.384953 1.17684
Number of obs: 73421, groups: s, 2972; d, 1128; dept:service, 28

Fixed effects:
            Estimate Std. Error t value
(Intercept)  3.25521    0.02824   115.3
```

(Fitting this complex model to a moderately large data set takes less than two minutes on a modest laptop computer purchased in 2006. Although this is more time than required for earlier model fits, it is a remarkably short time for fitting a model of this size and complexity. In some ways it is remarkable that such a model can be fit at all on such a computer.)

All three estimated standard deviations of the random effects are less than $\widehat{\sigma}$, with $\widehat{\sigma}_3$, the estimated standard deviation of the random effects for the `dept:service` interaction, less than one-tenth the estimated residual standard deviation.

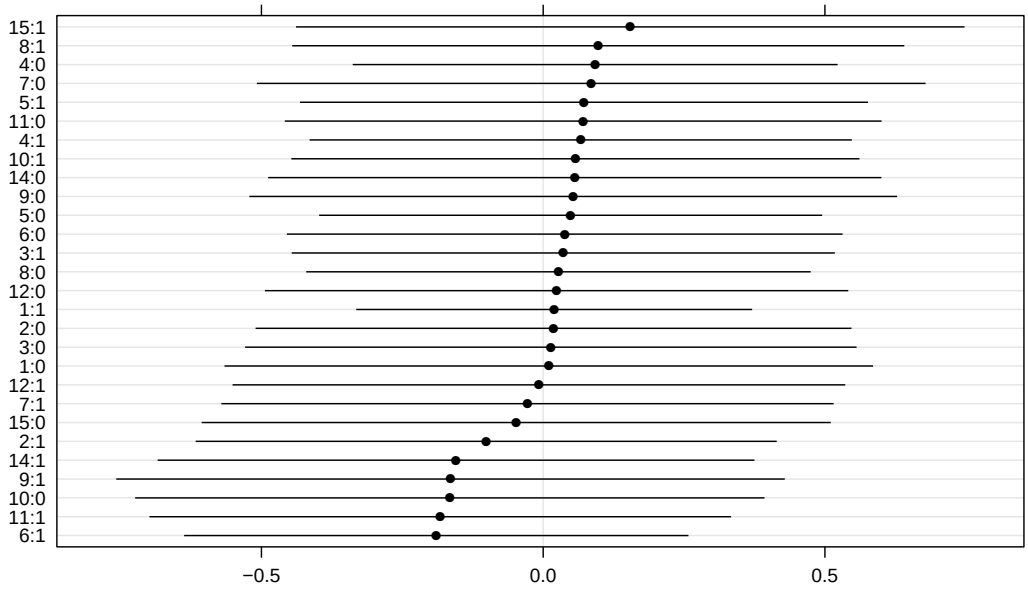


Fig. 2.14 95% prediction intervals on the random effects for the `dept:service` factor in model `fm4` fit to the `InstEval` data.

It is not surprising that zero is within all of the prediction intervals on the random effects for this factor (Fig. 2.14). In fact, zero is close to the middle of all these prediction intervals. However, the p-value for the LRT of $H_0 : \sigma_3 = 0$ versus $H_a : \sigma_3 > 0$

```
> fm4a <- lmer(y ~ 1 + (1|s) + (1|d), InstEval, REML=0)

> anova(fm4a,fm4)

Data: InstEval
Models:
fm4a: y ~ 1 + (1 | s) + (1 | d)
fm4: y ~ 1 + (1 | s) + (1 | d) + (1 | dept:service)
     Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
fm4a  4 237786 237823 -118889
fm4   5 237663 237709 -118827 124.43      1  < 2.2e-16
```

is highly significant. That is, we have very strong evidence that we should reject H_0 in favor of H_a.

The seeming inconsistency of these conclusions is due to the large sample size ($n = 73421$). When a model is fit to a very large sample even the most subtle of differences can be highly "statistically significant". The researcher or data analyst must then decide if these terms have practical significance, in addition to the apparent statistical significance.

The large sample size also helps to assure that the parameters have good normal approximations. We could profile this model fit but doing so would take a very long time and, in this particular case, the analysts are more

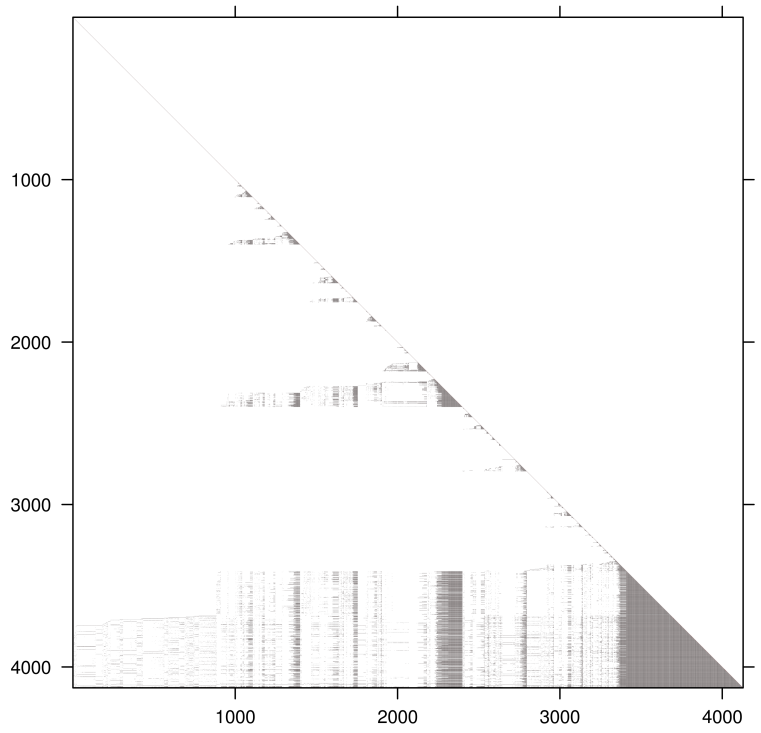


Fig. 2.15 Image of the sparse Cholesky factor, **L**, from model `fm4`

interested in a model that uses fixed-effects parameters for the instructors, which we will describe in the next chapter.

Before leaving this model we examine the sparse Cholesky factor, **L**, (Fig. 2.15), which is of size 4128×4128. Even as a sparse matrix this factor requires a considerable amount of memory,

```
> object.size(env(fm4)$L)

6904640 bytes

> unclass(round(object.size(env(fm4)$L)/2^20, 3))  # size in megabytes

[1] 6.585
```

but as a triangular dense matrix it would require nearly 10 times as much. There are $(4128 \times 4129)/2$ elements on and below the diagonal, each of which would require 8 bytes of storage. A packed lower triangular array would require

```
> (8 * (4128 * 4129)/2)/2^20   # size in megabytes

[1] 65.01965
```

megabytes. The more commonly used full rectangular storage requires

```
> (8 * 4128^2)/2^20   # size in megabytes
```

```
[1] 130.0078
```

megabytes of storage.

The number of nonzero elements in this matrix that must be updated for each evaluation of the deviance is

```
> nnzero(as(env(fm4)$L, "sparseMatrix"))
```

```
[1] 566960
```

Comparing this to 8522256, the number of elements that must be updated in a dense Cholesky factor, we can see why the sparse Cholesky factor provides a much more efficient evaluation of the profiled deviance function.

2.4 Chapter Review

A simple, scalar random effects term in an `lmer` model formula is of the form `(1|fac)`, where `fac` is an expression whose value is the *grouping factor* of the set of random effects generated by this term. Typically, `fac` is simply the name of a factor, such as in the terms `(1|sample)` or `(1|plate)` in the examples in this chapter. However, the grouping factor can be the value of an expression, such as `(1|dept:service)` in the last example.

Because simple, scalar random-effects terms can differ only in the description of the grouping factor we refer to configurations such as crossed or nested as applying to the terms although it is more accurate to refer to the configuration as applying to the grouping factors.

A model formula can include several such random effects terms. Because configurations such as nested or crossed or partially crossed grouping factors are a property of the data, the specification in the model formula does not depend on the configuration. We simply include multiple terms.

One apparent exception to this rule occurs with implicitly nested factors, in which the levels of one factor are only meaningful within a particular level of the other factor. In the `Pastes` data, levels of the `cask` factor are only meaningful within a particular level of the `batch` factor. A model formula of

```
strength ~ 1 + (1 | cask) + (1 | batch)
```

would result in a fitted model that did not appropriately reflect the sources of variability in the data. Following the simple rule that the factor should be defined so that distinct experimental or observational units correspond to distinct levels of the factor will avoid such ambiguity.

For convenience, a model with multiple, nested random-effects terms can be specified as

```
strength ~ 1 + (1 | batch/cask)
```

which internally is re-expressed as

```
strength ~ 1 + (1 | batch) + (1 | batch:cask)
```

We will avoid terms of the form `(1|batch/cask)`, preferring instead an explicit specification with simple, scalar terms based on unambiguous grouping factors.

The `InstEval` data, described in Sec. 2.3.1, illustrate some of the characteristics of the real data to which mixed-effects models are now fit. There is a large number of observations associated with several grouping factors; two of which, student and instructor, have a large number of levels and are partially crossed. Such data are common in sociological and educational studies but until now it has been very difficult to fit models that appropriately reflect such a structure. Much of the literature on mixed-effects models leaves the impression that multiple random effects terms can only be associated with nested grouping factors. The emphasis on hierarchical or multilevel configurations is an artifact of the computational methods used to fit the models, not the models themselves.

The parameters of the models fit to small data sets have properties similar to those for the models in the previous chapter. That is, profile-based confidence intervals on the fixed-effects parameter, β_0, are symmetric about the estimate but overdispersed relative to those that would be calculated from a normal distribution and the logarithm of the residual standard deviation, $\log(\sigma)$, has a good normal approximation. Profile-based confidence intervals for the standard deviations of random effects (σ_1, σ_2, etc.) are symmetric on a logarithmic scale except for those that could be zero.

Another observation from the last example is that, for data sets with a large numbers of observations, a term in a model may be "statistically significant" even when its practical significance is negligible.

Chapter 3
Models for Longitudinal Data

Longitudinal data consist of repeated measurements on the same subject (or some other "experimental unit") taken over time. Generally we wish to characterize the time trends within subjects and between subjects. The data will always include the response, the time covariate and the indicator of the subject on which the measurement has been made. If other covariates are recorded, say whether the subject is in the treatment group or the control group, we may wish to relate the within- and between-subject trends to such covariates.

In this chapter we introduce graphical and statistical techniques for the analysis of longitudinal data by applying them to a simple example.

3.1 The sleepstudy data

Belenky et al. [2003] report on a study of the effects of sleep deprivation on reaction time for a number of subjects chosen from a population of long-distance truck drivers. These subjects were divided into groups that were allowed only a limited amount of sleep each night. We consider here the group of 18 subjects who were restricted to three hours of sleep per night for the first ten days of the trial. Each subject's reaction time was measured several times on each day of the trial.

```
> str(sleepstudy)

'data.frame':        180 obs. of  3 variables:
 $ Reaction: num  250 259 251 321 357 ...
 $ Days    : num  0 1 2 3 4 5 6 7 8 9 ...
 $ Subject : Factor w/ 18 levels "308","309","310",..: 1 1 1 1 1 1 1 1 1 1 ...
```

In this data frame, the response variable `Reaction`, is the average of the reaction time measurements on a given subject for a given day. The two covariates are `Days`, the number of days of sleep deprivation, and `Subject`, the identifier of the subject on which the observation was made.

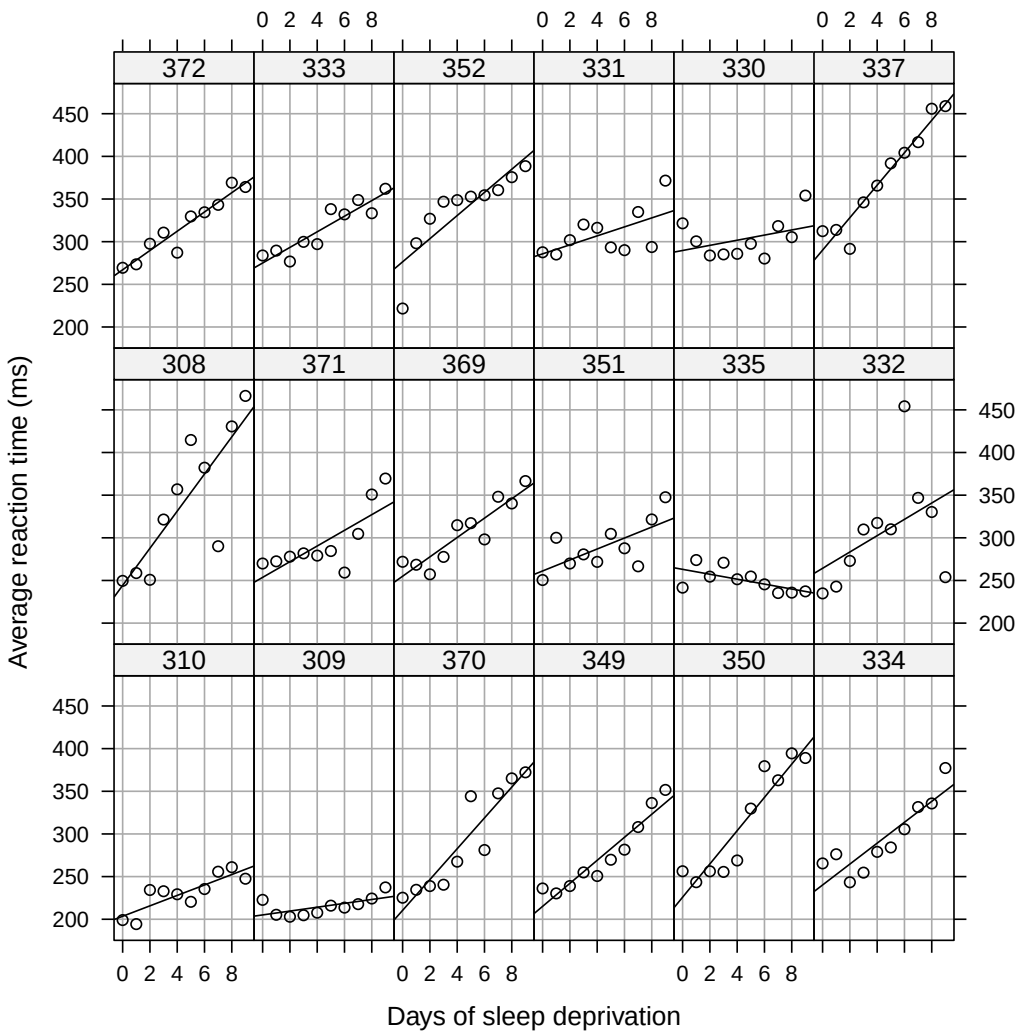


Fig. 3.1 A lattice plot of the average reaction time versus number of days of sleep deprivation by subject for the `sleepstudy` data. Each subject's data are shown in a separate panel, along with a simple linear regression line fit to the data in that panel. The panels are ordered, from left to right 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.

As recommended for any statistical analysis, we begin by plotting the data. The most important relationship to plot for longitudinal data on multiple subjects is the trend of the response over time by subject, as shown in Fig. 3.1. This plot, in which the data for different subjects are shown in separate panels with the axes held constant for all the panels, allows for examination of the time-trends within subjects and for comparison of these patterns between subjects. Through the use of small panels in a repeating pattern Fig. 3.1 conveys a great deal of information, the individual time trends for 18 subjects over 10 days — a total of 180 points — without being overly cluttered.

3.1.1 Characteristics of the Data Plot

The principles of "Trellis graphics", developed by Bill Cleveland and his coworkers at Bell Labs and implemented in the `lattice` package for R by Deepayan Sarkar, have been incorporated in this plot. As stated above, all the panels have the same vertical and horizontal scales, allowing us to evaluate the pattern over time for each subject and also to compare patterns between subjects. The line drawn in each panel is a simple least squares line fit to the data in that panel only. It is provided to enhance our ability to discern patterns in both the slope (the typical change in reaction time per day of sleep deprivation for that particular subject) and the intercept (the average response time for the subject when on their usual sleep pattern).

The aspect ratio of the panels (ratio of the height to the width) has been chosen, according to an algorithm described in Cleveland [1993], to facilitate comparison of slopes. The effect of choosing the aspect ratio in this way is to have the slopes of the lines on the page distributed around $\pm 45^\circ$, thereby making it easier to detect systematic changes in slopes.

The panels have been ordered (from left to right starting at the bottom row) by increasing intercept. Because the subject identifiers, shown in the strip above each panel, are unrelated to the response it would not be helpful to use the default ordering of the panels, which is by increasing subject number. If we did so our perception of patterns in the data would be confused by the, essentially random, ordering of the panels. Instead we use a characteristic of the data to determine the ordering of the panels, thereby enhancing our ability to compare across panels. For example, a question of interest to the experimenters is whether a subject's rate of change in reaction time is related to the subject's initial reaction time. If this is the case then we would expect that the slopes would show an increasing trend (or, less likely, a decreasing trend) in the left to right, bottom to top ordering.

There is little evidence in Fig. 3.1 of such a systematic relationship between the subject's initial reaction time and their rate of change in reaction time per day of sleep deprivation. We do see that for all the subjects, except 335, reaction time increases, more-or-less linearly, with days of sleep deprivation. However, there is considerable variation both in the initial reaction time and in the daily rate of increase in reaction time. We can also see that these data are balanced, both with respect to the number of observations on each subject, and with respect to the times at which these observations were taken. This can be confirmed with a cross-tabulation of `Subject` by `Days`.

```
> xtabs(~ Subject + Days, sleepstudy)

       Days
Subject 0 1 2 3 4 5 6 7 8 9
    308 1 1 1 1 1 1 1 1 1 1
    309 1 1 1 1 1 1 1 1 1 1
    310 1 1 1 1 1 1 1 1 1 1
    330 1 1 1 1 1 1 1 1 1 1
```

```
    331 1 1 1 1 1 1 1 1 1 1
    332 1 1 1 1 1 1 1 1 1 1
    333 1 1 1 1 1 1 1 1 1 1
    334 1 1 1 1 1 1 1 1 1 1
    335 1 1 1 1 1 1 1 1 1 1
    337 1 1 1 1 1 1 1 1 1 1
    349 1 1 1 1 1 1 1 1 1 1
    350 1 1 1 1 1 1 1 1 1 1
    351 1 1 1 1 1 1 1 1 1 1
    352 1 1 1 1 1 1 1 1 1 1
    369 1 1 1 1 1 1 1 1 1 1
    370 1 1 1 1 1 1 1 1 1 1
    371 1 1 1 1 1 1 1 1 1 1
    372 1 1 1 1 1 1 1 1 1 1
```

In cases like this where there are several observations (10) per subject and a relatively simple within-subject pattern (more-or-less linear) we may want to examine coefficients from within-subject fixed-effects fits. However, because the subjects constitute a sample from the population of interest and we wish to drawn conclusions about typical patterns in the population and the subject-to-subject variablity of these patterns, we will eventually want to fit mixed models so we begin doing so.

3.2 Mixed-effects models for the sleep data

Based on our preliminary graphical and analytical exploration of these data, we fit a mixed-effects model with two fixed-effects parameters, the intercept and slope of the linear time trend for the population, and two random effects for each subject. The random effects for a particular subject are the deviations in intercept and slope of that subject's time trend from the population values.

```
> (fm8 <- lmer(Reaction ~ Days + (Days|Subject), sleepstudy))

Linear mixed model fit by REML
Formula: Reaction ~ Days + (Days | Subject)
   Data: sleepstudy
REML
1744

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 612.091  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
```

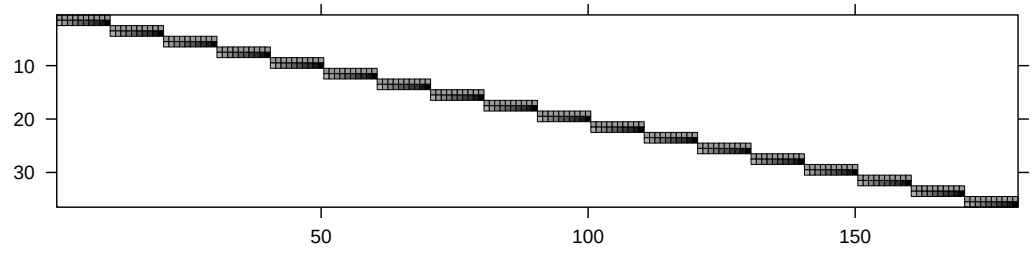


Fig. 3.2 Image of $\mathbf{Z}^\mathsf{T}$ for model `fm8`

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

We see that this model incorporates both and intercept and a slope (w.r.t. `Days`) in both the fixed effects and the random effects. In linear model formulas the intercept term is implicit. If we had written the formula as

```
Reaction ~ 1 + Days + (1 + Days | Subject)
```

the fitted model would be the same. Many people prefer to include the intercept explicitly in the formula to emphasize the relationship between terms in the formula and coefficients or random effects in the model. Others omit these implicit terms to economize on the amount of typing required.

As this is the first time we have seen a random effects term other than for a simple, scalar random effect we should discuss the general form of such terms. As previously described, the expression on the right hand side of the vertical bar is evaluated as a factor, the *grouping factor* for the term. The expression on the left hand side is evaluated as a linear model formula, producing a model matrix. Because the left hand side of `(1 + Days|Subject)` produces a model matrix with two columns, there will be two random effects associated with each level of `Subject`. That is, there is a total of 36 random effects generated by this term. We say that we have a *vector-valued random effect*, an intercept effect and a slope effect, for each level of the `Subject` factor.

As shown in Fig. reffig:fm8Zt, the 36 columns of $\mathbf{Z}$ (rows of $\mathbf{Z}^\mathsf{T}$ in the figure) can be regarded as 18 pairs, one pair for each level of `Subject`. The first column in each pair is the indicator for that level of `Subject`. The second column in each pair is the value of the `Days` variable but only for the measurements on that subject.

One way to imagine the process of generating $\mathbf{Z}$ is to start with the 180×2 model matrix for the linear model formula `1 + Days` and the 180×18 indicator matrix for the levels of `Subject`.

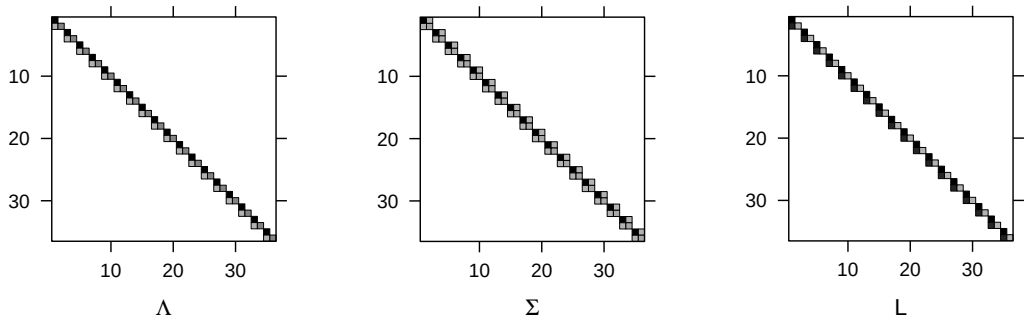


Fig. 3.3 Images of Λ, Σ and $\mathbf{L}$ for model `fm8`

The images of Λ, Σ and L for this model (Fig. 3.3) show 18 triangular blocks of size 2 along the diagonal of Λ, generating 18 square, symmetric blocks of size 2 along the diagonal of Σ. The 18 symmetric blocks on the diagonal of Σ are identical, although this may not be obvious from the figure. Overall we estimate two standard deviations and a correlation for the vector-valued random effect of size 2, as shown in the model summary.

Often the variances and the covariance of random effects are quoted, rather than the standard deviations and the correlation shown here. We have already seen that the variance of a random effect is a poor scale on which to quote the estimate because confidence intervals on the variance are so badly skewed. It is more sensible to assess the estimates of the standard deviations of random effects or possibly the logarithms of the standard deviations, when we can be confident that 0 is outside the region of interest. We do display the estimates of the variances of the random effects but mostly so that the user can compare these estimates to those from other software or for cases where an estimated of a variance is expected (sometimes even required) to be given in reporting a model fit.

We do not quote estimates of covariances of vector-valued random effects because the covariance is a difficult scale to interpret. We know that a correlation must be between -1 and 1. A correlation estimate close to those extremes indicates that Σ is close to singular and the model is not well formulated.

To specify a model with independent random effects by subject for the intercept and the slope we will use two random-effects terms and write the formula as

```
Reaction ~ Days + (1 | Subject) + (0 + Days | Subject)
```

In this formula the random effects for the intercept and the random effects for the slope are modeled as independent random variables because they are declared in different expressions. Note that to specify a random intercept given subject we must explicitly include the intercept term `1` in `(1|Subject)` because there are no other terms. Similarly in the second expression we sup-

press the implicit intercept term by using `(0+Days|Subject)`, read as "no intercept and `Days` by `Subject`". An alternative expression for `Days` without an intercept by `Subject` is `(Days - 1 | Subject)`.

We delay further discussion on the mathematical form of the model and the interpretation of the model formulae until §**??**. At this point let us fit these models and examine the parameter estimates.

We can fit the first model, store the result as `fm8` (sleepstudy model 1), and ask for a brief display of the results with

```
> (fm8 <- lmer(Reaction ~ Days + (Days|Subject), sleepstudy))

Linear mixed model fit by REML
Formula: Reaction ~ Days + (Days | Subject)
   Data: sleepstudy
REML
1744

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 612.091  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
```

(The extra set of parentheses surrounding the assignment causes the fitted model to display itself. Normally the result of an assignment is not displayed.)

This brief display includes information on the criterion used to fit the model (restricted maximum likelihood or REML, see §**??** for a formal definiton), the model formula and the data to which it was fit, some information on the quality of the fit and information on the parameter estimates. For the moment we will concentrate on the parameter estimates.

The estimates of the fixed effects parameters are $\widehat{\beta} = (251.41, 10.467)^{\mathsf{T}}$. These represent a typical initial reaction time (i.e. without sleep deprivation) in the population of about 250 milliseconds, or 1/4 sec., and a typical increase in reaction time of a little more than 10 milliseconds per day of sleep deprivation.

The estimated variance-covariance matrix for the random effects is displayed by giving the variances of these random variables, the corresponding standard deviations and any estimated correlations. Note that the columns labeled `Variance` and `Std.Dev.` in this section are redundant in that each entry

in the `Std.Dev.` column is simply the square root of the corresponding variance estimate. These estimates are expressed in both the variance scale and the standard deviation scale because both are useful in interpretation. (Some readers may be tempted to interpret the elements of the `Std.Dev.` column as standard errors of the variance estimates. Don't do that. These are not standard errors.)

The estimated subject-to-subject variation in the intercept corresponds to a standard deviation of about 25 ms. A 95% prediction interval on this random variable would be approximately ±50 ms. Combining this range with a population estimated intercept of 250 ms. indicates that we should not be surprised by intercepts as low as 200 ms. or as high as 300 ms. This range is consistent with the reference lines shown in Figure 3.1 and the intervals shown in Figure **??**.

Similarly, the estimated subject-to-subject variation in the slope corresponds to a standard deviation of about 6 ms./day so we would not be surprised by slopes as low as $10.5 - 2 \cdot 6 = -1.5$ ms./day or as high as $10.5 + 2 \cdot 6 = 22.5$ ms./day. Again, the conclusions from these rough, "back of the envelope" calculations are consistent with our observations from Figures 3.1 and **??**.

The estimated residual standard deviation is about 25 ms. leading us to expect a scatter around the fitted lines for each subject of up to ±50 ms. From Figure 3.1 we can see that some subjects (309, 372 and 337) appear to have less variation than ±50 ms. about their within-subject fit but others (308, 332 and 331) may have more.

Finally, we see the estimated within-subject correlation of the random effect for the intercept and the random effect for the slope is very low, 0.066, confirming our impression that there is little evidence of a systematic relationship between these quantities. In other words, observing a subject's initial reaction time does not give us much information for predicting whether their reaction time will be strongly affected by each day of sleep deprivation or not.

By fitting model `fm9` with independent random effects for intercept and slope and comparing this fitted model to `fm8` we can assess this claim using a statistical hypothesis test.

```
> (fm9 <- lmer(Reaction ~ Days + (1|Subject) + (0+Days|Subject), sleepstudy))

Linear mixed model fit by REML
Formula: Reaction ~ Days + (1 | Subject) + (0 + Days | Subject)
   Data: sleepstudy
REML
1744

Random effects:
 Groups   Name        Variance Std.Dev.
 Subject  (Intercept) 627.569  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      1.560    6.71

Correlation of Fixed Effects:
     (Intr)
Days -0.184

> anova(fm9, fm8)

Data: sleepstudy
Models:
fm9: Reaction ~ Days + (1 | Subject) + (0 + Days | Subject)
fm8: Reaction ~ Days + (Days | Subject)
    Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
fm9  5 1762.0 1778.0 -876.00
fm8  6 1763.9 1783.1 -875.97 0.0639      1     0.8004
```

We can see that the fitted model `fm9` is quite similar to `fm8` except for the obvious difference that there is no within-subject correlation of the random effects in `fm9`. The estimates of all the other parameters, which are common to the two models, are practically unchanged.

The call to `anova` compares the two fitted models using a likelihood ratio test, which evaluates the change in the quality of the fits, as measured by the deviance (defined in §**??**), relative to the change in the number of parameters. The results of this test indicate that the model `fm8` does not fit significantly better than the model `fm9` and hence we prefer the model `fm9` which has fewer parameters.

We conclude that there is significant variation between subjects in both the initial reaction time and in the rate of change in reaction time with respect to days of sleep deprivation but that these changes are not correlated. That is knowing a person's initial reaction time does not help us to predict their response to sleep deprivation.

3.3 Assessing the precision of the parameter estimates

Plots of the profile ζ (Figure 3.4) show that confidence intervals σ_1 and σ_2 will be slightly skewed; those for $\log(\sigma)$ will be symmetric and well-approximated by methods based on quantiles of the standard normal distribution and those for the fixed-effects parameters, β_1 and β_2 will be symmetric and slightly over-dispersed relative to the standard normal. For example, the 95% profile-based confidence intervals are

```
> confint(pr2)
```

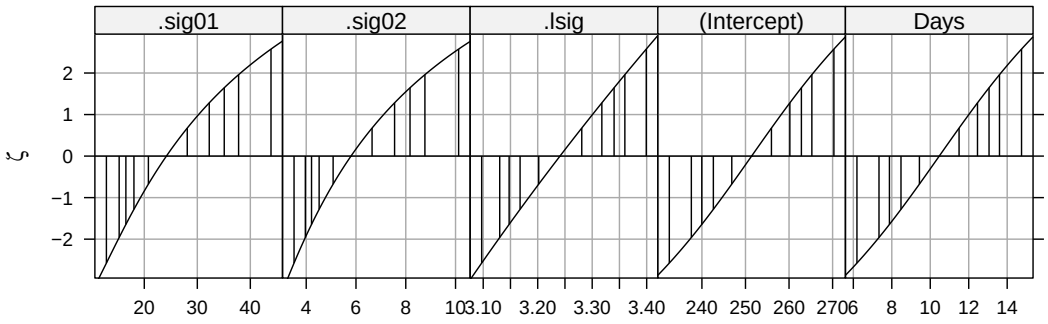


Fig. 3.4 Profile zeta plot for each of the parameters in model `fm9`. The vertical lines are the endpoints of 50%, 80%, 90%, 95% and 99% profile-based confidence intervals for each parameter.

```
                 2.5 %     97.5 %
.sig01       0.6335658  1.1821040
.sig02       1.0957822  3.5563194
.lsig       -0.7218645 -0.4629033
(Intercept) 21.2666274 24.6778176
```

The profile pairs plot (Figure 3.5)

In the previous section we used a likelihood ratio test to assess whether the covariance of the random effects within subject is significantly different from zero. This is an example of a statistical hypothesis test, which is one form of statistical inference. Another, related, form of statisticial inference is assessing the precision of the parameter estimates, say by forming confidence intervals or more general confidence regions.

For some statistical models it is possible to derive the theoretical distributions of the parameter estimates and use these theoretical distributions to create confidence intervals or regions. At present, the theoretical tools to analyze the general form of a linear mixed model in this "exact" approach are not available and confidence intervals and regions are typically created using approximations, especially what are called "asymptotic" approximations that are expected to perform well for large data sets.

Recently another approach to statistical inference, using Markov chain Monte Carlo (MCMC) samples from the (Bayesian) posterior distribution of the parameters, has gained popularity. Although MCMC is a computationally intensive approach to inference the current availability of inexpensive, powerful computers has made it feasible for models such as linear mixed models.

An advantage of using MCMC samples to assess the precision of the parameter estimates is that this approach uses the actual distribution of the parameters and not an approximation. Furthermore we can use graphical

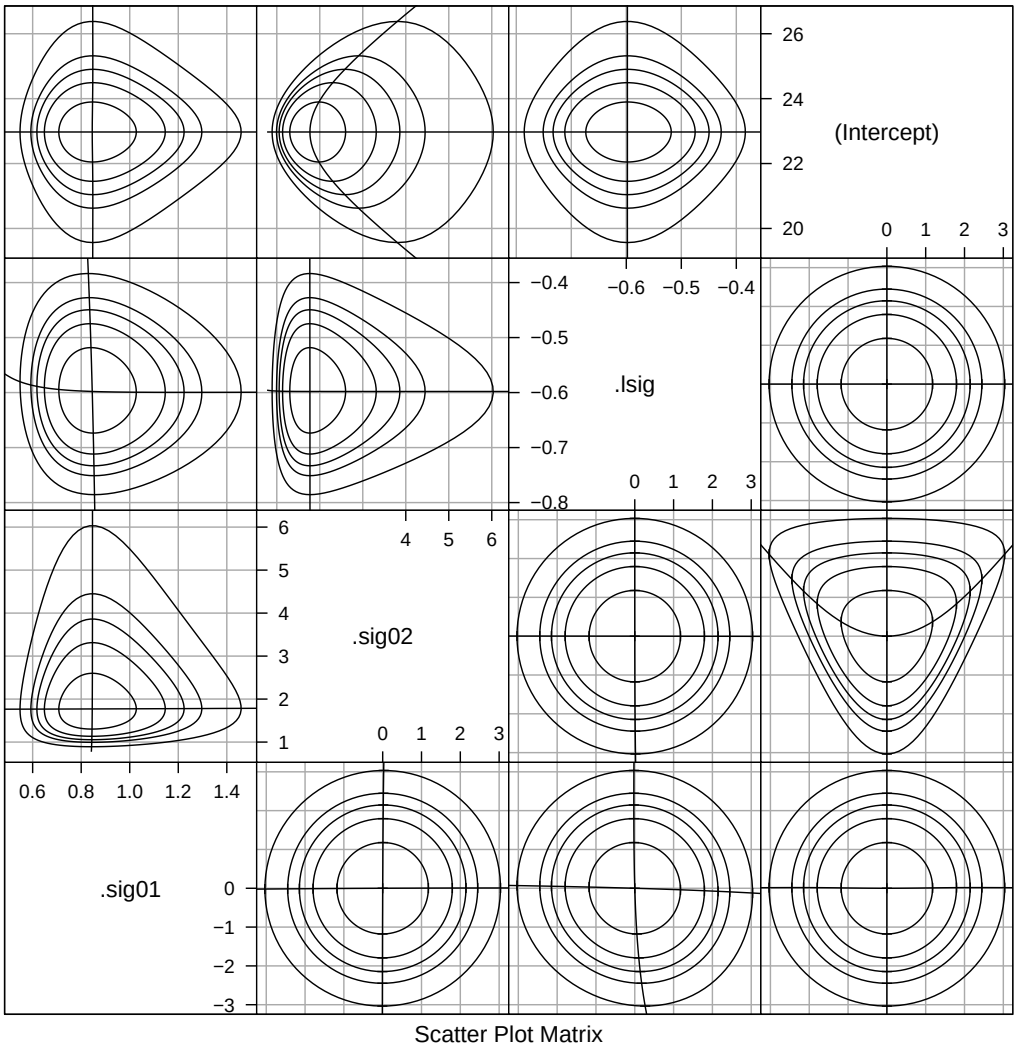


Fig. 3.5 Profile pairs plot for the parameters in model `fm9`. The contour lines correspond to marginal 50%, 80%, 90%, 95% and 99% confidence regions based on the likelihood ratio. Panels below the diagonal represent the (ζ_i, ζ_j) parameters; those above the diagonal represent the original parameters.

techniques to visualize these distributions and provide insight into the behavior of the parameters in the model.

MCMC sampling methods are based on a Bayesian formulation of the linear mixed model in which the parameters are considered to be random variables that have a *prior* distribution (prior in the sense of "before the data are known") and a *posterior*, or "after the data are known" distribution. Instead of confidence intervals on the parameters we will formulate *highest posterior density* (HPD) intervals [Box and Tiao, 1973]. An 95% HPD interval on a parameter is the shortest interval that contains 95% of the probability content of the posterior distribution. It is the Bayesian equivalent of a 95% confidence interval on the parameter.

Details of the particular Bayesian formulation of the linear mixed model that we use, including the choice of prior distributions for the parameters, are given in §**??**.

3.3.1 Posterior distributions from model `fm9`

The function `mcmcsamp` applied to a fitted lmer model produces an MCMC sample from the posterior distribution of the parameter estimates, from which we can evaluate HPD intervals. Let us create and store a sample of size 10,000 from the posterior distribution of the parameters in model `fm9`.

The `HPDinterval` function creates HPD intervals on each of the parameters in the sample. By default it returns intervals whose empirical probability content is 95% (this can be changed, if desired).

Notice that in the intervals all the variance parameters are reported on the logarithm scale. The reason for taking this transformation is because the logarithm of a variance tends to be symmetrically distributed as shown by the density plots in Figure **??**

Not only are the posterior distributions on this scale symmetric, they are very close to normal distributions as shown by their normal probability plots (Figure **??**)

3.4 Examining the random effects

- Although the random effects **b** behave like parameters in the linear predictor, technically they are not parameters in the model.
- Instead of referring to "estimates" of the random effects it is customary to refer to "predictors" - in particular, the best linear unbiased predictors or BLUPs.
- These values are also the modes of the conditional distribution (i.e. given the data **y** and the estimates of β, σ^2 and Σ) of **b**.
- For linear mixed model the conditional distribution $[\mathbf{b}|\mathbf{y}, \sigma^2, \Sigma]$ is normal (Gaussian) hence the modes are also the conditional means.
- The `ranef` extractor function returns these conditional modes evaluated at the parameter estimates.

```
> (rr1 <- ranef(fm8))

$Subject
    (Intercept)        Days
308   2.2585637   9.1989719
309 -40.3985926  -8.6197003
310 -38.9602618  -5.4488771
330  23.6905107  -4.8143332
```

```
331  22.2602137  -3.0698963
332   9.0395301  -0.2721713
333  16.8404388  -0.2236257
334  -7.2325828   1.0745766
335  -0.3336928 -10.7521593
337  34.8903640   8.6282815
349 -25.2101221   1.1734162
350 -13.0699646   6.6142061
351   4.5778381  -3.0152576
352  20.8636008   3.5360118
369   3.2754542   0.8722164
370 -25.6128825   4.8224666
371   0.8070404  -0.9881552
372  12.3145445   1.2840288
```

- For this model we can combine the BLUPs of the random effects and the estimates of the fixed effects to get BLUPs for the within-subject coefficients.
- These BLUPs will be "shrunken" towards the fixed-effects estimates relative to the estimated coefficients from only that subjects data. John Tukey called this "borrowing strength" between subjects.
- Plotting the shrinkage of the within-subject coefficients shows that some of the coefficients are considerably shrunken toward the fixed-effects estimates.
- However, comparing the within-group and mixed model fitted lines shows that large changes in coefficients occur in the noisy data. Precisely estimated within-group coefficients are not changed substantially.

3.4.1 Prediction intervals on the random effects

- For the linear mixed model we can calculate both the means and the variances of the random-effects conditional on the estimated values of the model parameters, which allows us to calculate prediction intervals on the values of individual random effects.
- We plot the prediction intervals as a normal probabity plot so we can see the overall shape of the distribution of the means and which of the random effects are "significantly different" from zero.
- Note that failure of the conditional means of the random effects to look like a normal (Gaussian) distribution is not terribly alarming. It is the "prior" distribution of the random effects that is assumed to be normal. The conditional means or BLUPs are strongly influenced by the data and may appear non-normal.

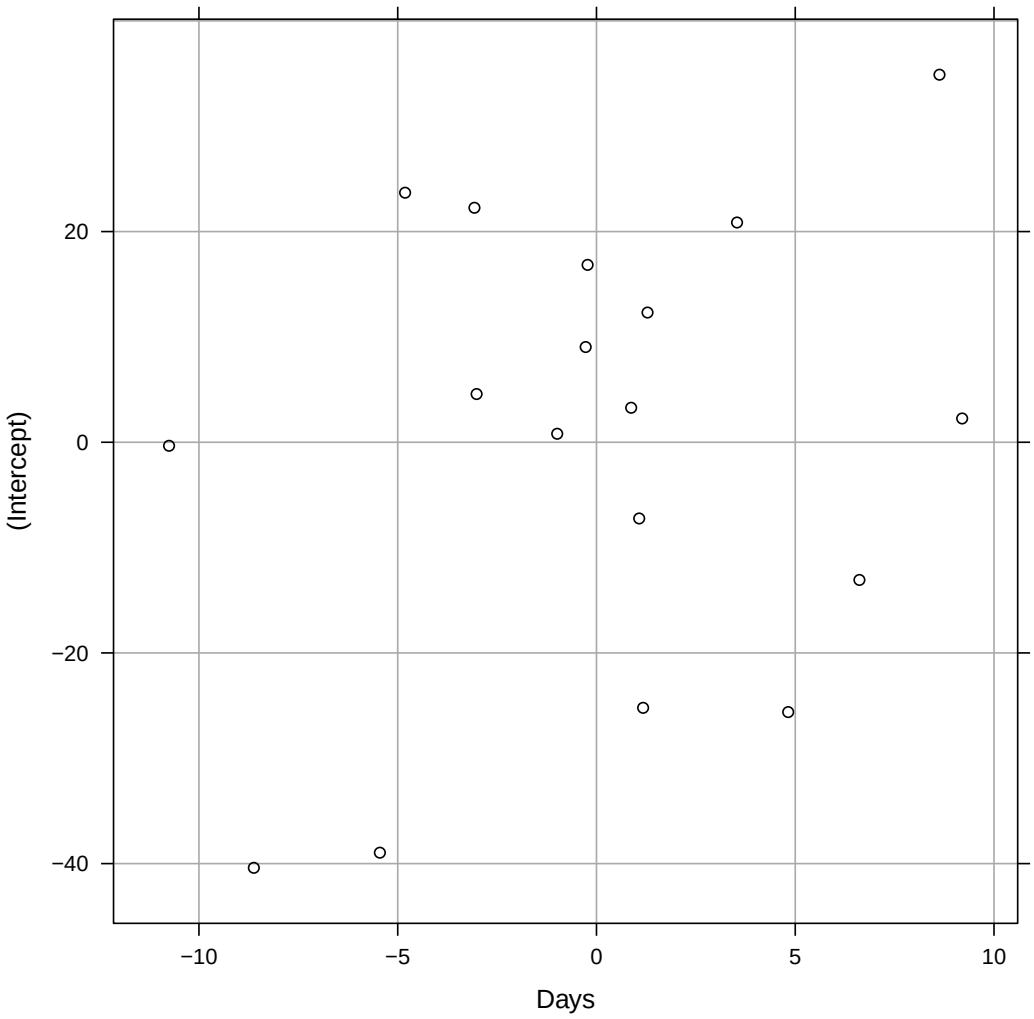


Fig. 3.6 Scatterplot of the conditional modes, or BLUPs, of the random effects for model `Fm8`. Each point represents the mode of the distribution of the random effects for the intercept and slope associated with one of the subjects.

3.5 Model specification for `lmer`

A linear mixed-effects model to be fit by `lmer` is specified by the `formula` argument. For model `fm8` the formula is

```
Reaction ~ Days + (Days | Subject)
```

which can be read as "`Reaction` is modeled by `Days` and `Days` given `Subject`". That is, the response, which is the variable named `Reaction`, is to be modeled by one conditional term, `(Days|Subject)`, and one unconditional term, `Days`.

A conditonal term (any term including the vertical bar, `|`) contributes to the random effects specification. An unconditional term contributes to the fixed effects specification.

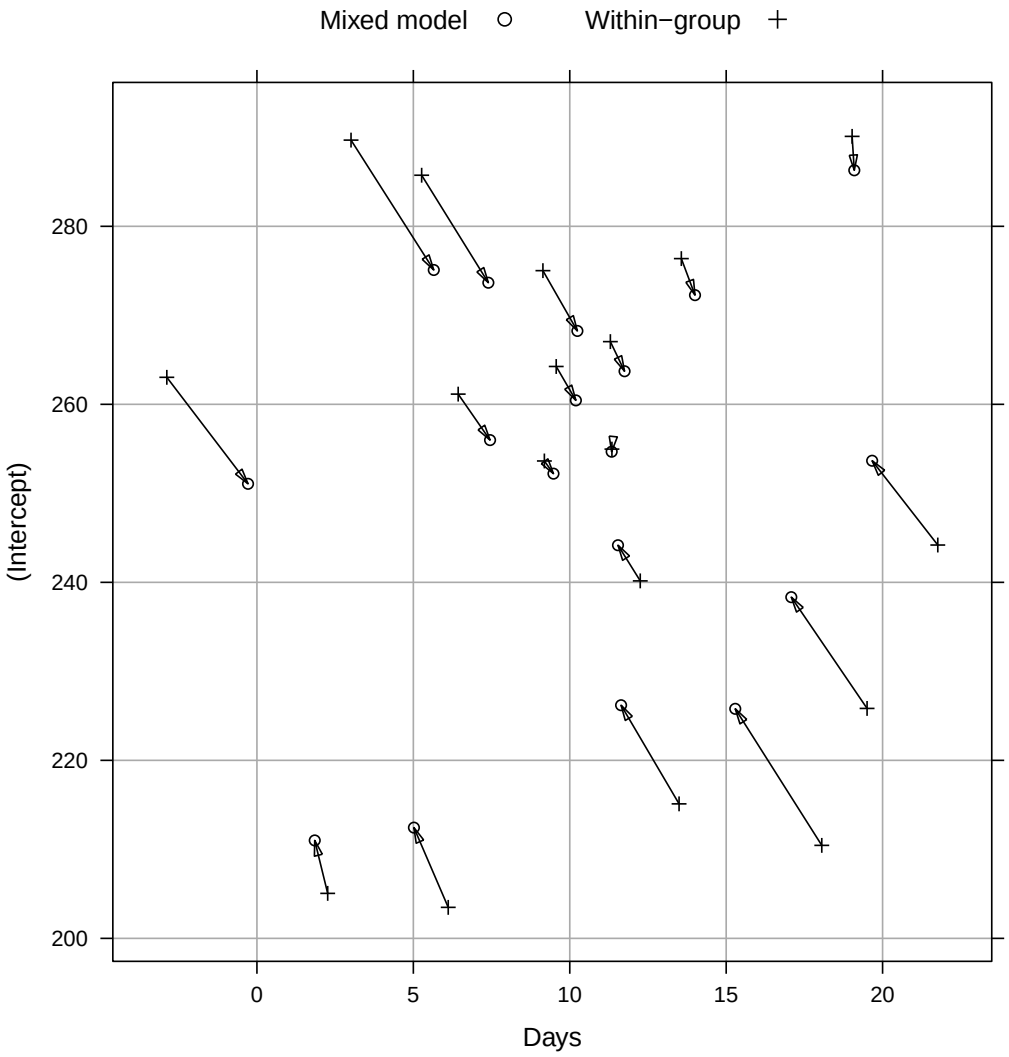


Fig. 3.7 Comparison of the within-subject estimates of the intercept and slope for each subject and the conditional modes of the per-subject intercept and slope. Each pair of points joined by an arrow are the within-subject and conditional mode estimates for the same subject. The arrow points from the within-subject estimate to the conditional mode for the mixed-effects model.

The unconditional terms, together with the data to be fit, generate an $n \times p$ fixed-effects model matrix $\mathbf{X}$ according to the rules that we describe in §**??**. The dimensions n and p are the number of observations and the dimension of the fixed-effects parameter vector $\boldsymbol{\beta}$, respectively. In this case of model `fm8` the only unconditional term, `Days`, is the name of a numeric variable, which is incorporated as a column, labelled `Days` in $\mathbf{X}$. By convention, the intercept term, which generates a column of 1's labelled `(Intercept)`, is included implicitly in the model specification. (This column can be suppressed if desired, as shown below.)

The first few rows of the 180×2 model matrix $\mathbf{X}$ for model `fm8` are

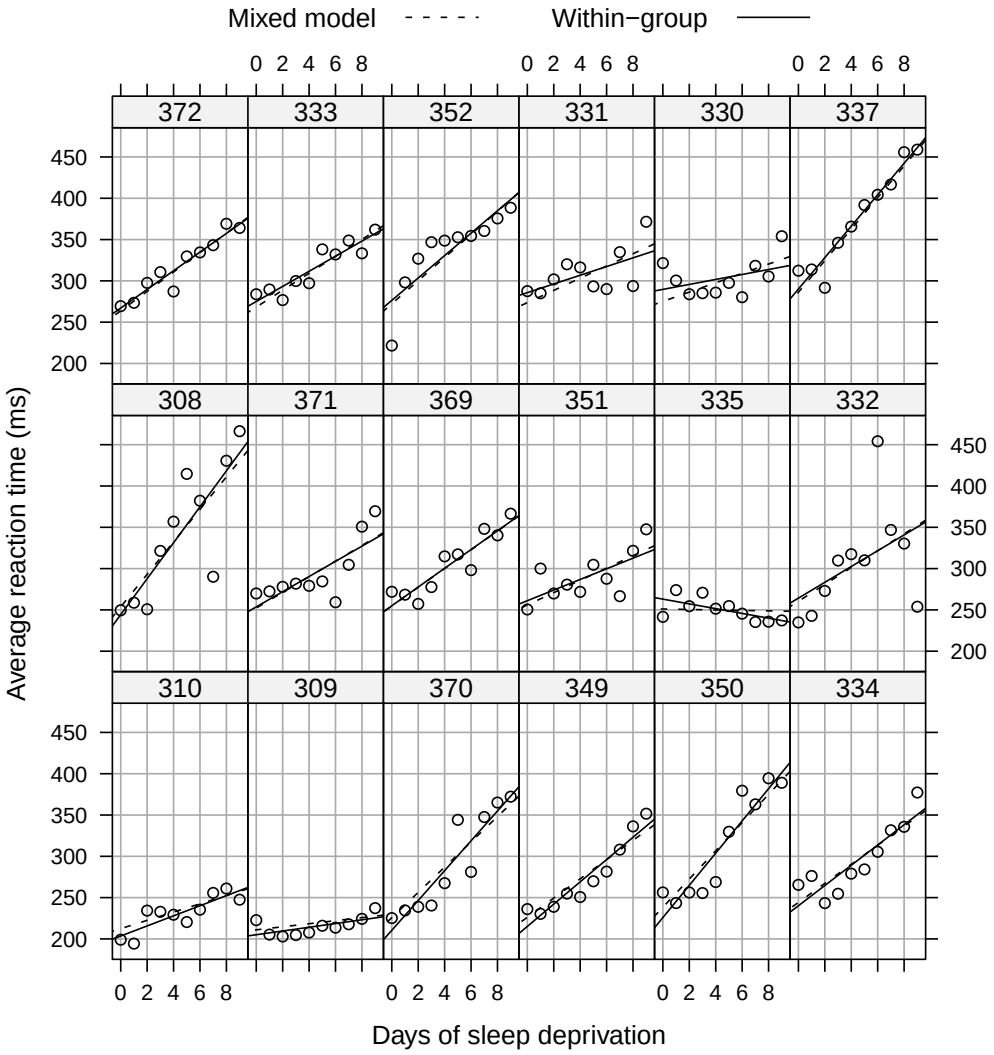


Fig. 3.8 Comparison of the predictions from the within-subject fits with those from the conditional modes of the subject-specific parameters in the mixed-effects model.

Fig. 3.9 Prediction intervals on the random effects per subject.

```
> head(model.matrix(fm8))

6 x 2 Matrix of class "dgeMatrix"
     (Intercept) Days
[1,]           1    0
[2,]           1    1
[3,]           1    2
[4,]           1    3
[5,]           1    4
[6,]           1    5
```

Each conditional term in the model formula generates a set of random effects and variance-covariance matrix for these random effects. In a condi-

tional term the expression on the right hand side of the `|` is evaluated as a factor called the *grouping factor* for the term. Because a factor associates one of a finite set of levels with each observation, we can consider a factor as dividing the observations into groups corresponding to the levels of the factor. Observations within such groups share a set of random effects. The expression on the left of the `|` in a conditional term is evaluated as a linear model formula and determines the number and form of the random effects associated with each level of the grouping factor. Thus `(Days|Subject)` designates an (implicit) intercept coefficient and a `Days` coefficient for each level of the `Subject` grouping factor producing, as we have seen, 36 random effects — two for each of the 18 subjects.

Random effects generated by different conditional terms are independent, as are random effects corresponding to different levels of the grouping factor in the same conditional term.

Each group of random effects models some of the variation in the response. There is one further level of variation in the model - the "per-observation" or "residual" noise. It is the unexplained variation or what is "left over" after we have modeled all the other sources of variation in the model. This level of variation is modeled as an n-dimensional vector $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ where $\mathbf{I}$ is an identity matrix. That is, the elements of $\boldsymbol{\varepsilon}$ is independent and identical normal random variates with mean zero and variance σ^2.

The general model allows for multiple conditional terms in a model specification generating multiple groups of random effects. Let k be the number of conditional terms, $n_i, i = 1, \ldots, k$ be the number of levels of the grouping factor for the ith such term and q_i be the number of random effects associated with each level of the grouping factor.

If we represent all the random effects as a vector $\mathbf{b}$, of length $q = \sum_{i=1}^{k} q_i n_i$, we can write the model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}, \quad \mathbf{b} \sim \mathcal{N}(\mathbf{0}, \Sigma), \boldsymbol{\varepsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right), \mathbf{b} \perp \boldsymbol{\varepsilon} \tag{3.1}$$

where $\mathbf{Z}$ is an $n \times q$ model matrix generated from the conditional terms, Σ is a $q \times q$ variance-covariance matrix, also generated from the conditional terms, and the symbol $\perp$ denotes independent random variables.

Although the total number of random effects, q, can be large and hence the dimensions of $\mathbf{Z}$ and Σ in the general form (3.1) can be very large, these matrices are sparse and patterned and thus are determined by a relatively small number of values.

For model `fm8`, $k = 1$, $q_1 = 2$ and $n_i = 18$ so, as we have seen, $q = 36$. However, the assumptions of independence of random effects associated with different subjects means that the 36×36 matrix Σ consists of the 2×2 matrix Σ_1 repeated 18 times in the pattern

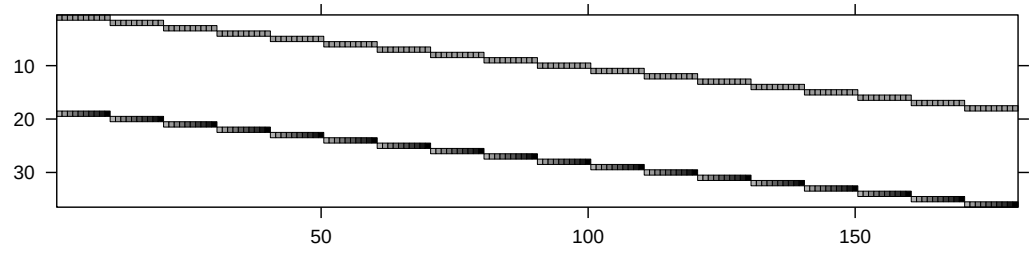


Fig. 3.10 Image of $\mathbf{Z}^\mathsf{T}$, the transpose of $\mathbf{Z}$, the random effects model matrix in model `fm9`

$$\Sigma = \begin{bmatrix} \Sigma_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \Sigma_1 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \Sigma_1 \end{bmatrix} \tag{3.2}$$

Also, although the model matrix $\mathbf{Z}$ has dimension 180×36, all but two of the elements in any given row are known to be zero. Each row corresponds to one and only one subject and the model for that row incorporates only the two random effects associated with that subject. The other 34 random effects associated with other subjects are multiplied by zero.

The matrix $\mathbf{Z}$ is evaluated and stored as a sparse matrix. The first few rows of $\mathbf{Z}$ for model `fm8` are

```
6 x 36 sparse Matrix of class "dgCMatrix"

[1,] 1 0 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
[2,] 1 1 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
[3,] 1 2 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
[4,] 1 3 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
[5,] 1 4 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
[6,] 1 5 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
```

These rows correspond to the first 6 observations on the first subject. In the representation as a sparse matrix an element that is known to be zero prints as a '.'. (The (1,2) element of this matrix does have the value zero because the value of the `Days` variable is zero for this observation. It could have another value if, for example, we renumbered the `Days` and thus is not a systematic zero in the matrix.)

Easier to understand, perhaps, is the image of $\mathbf{Z}^\mathsf{T}$ in Figure 3.10.

In the model $\mathbf{b}$ is a random variable

$$\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \Sigma) \tag{3.3}$$

and it is the elements of Σ that we estimate. Because we assume that random effects associated with different grouping factors are indep which we can write as If we write the model matrix for the entire random effects vector as $\mathbf{Z}$ and the variance-cov matrix, which we call $\mathbf{Z}$, for the random effects vector.

contribute to another model matrix, which we call $\mathbf{Z}$, through a slightly more complicated mechanism. In a conditional term the expression on the left of the `|` is interpreted as a linear model formula and used to create a model matrix while the expression on the right of the `|` is evaluated as a factor, called the *grouping factor* for the term. We allow multiple conditional terms in a model specification so we will refer to the ith conditional term even though there is only one such conditional term in the model specification for model `fm8`.

If the model matrix from the expression on the left of the ith conditional has q_i columns and the grouping factor has n_i levels then the expression contributes $q_i n_i$ columns to the matrix $\mathbf{Z}$.

3.6 Conclusions from the example

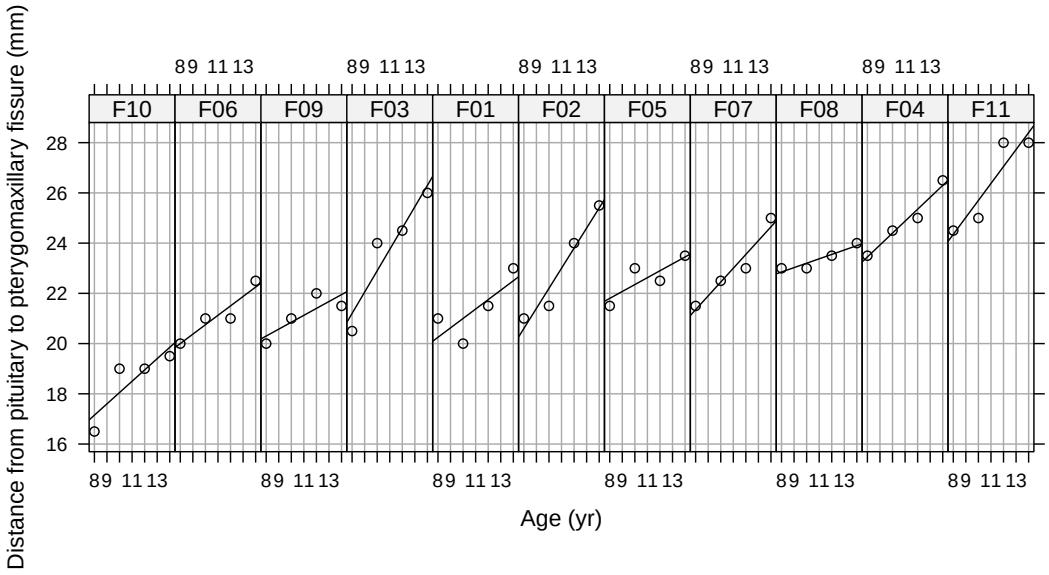
- Carefully plotting the data is enormously helpful in formulating the model.
- It is relatively easy to fit and evaluate models to data like these, from a balanced designed experiment.
- For a linear mixed model the estimates of the fixed effects typically have a symmetric distribution close to a Gaussian distribution.
- The distribution of the variance components or the covariances are not symmetric, which is why we transform these parameters to a symmetric scale.
- We use the MCMC sample to create confidence (actually HPD) intervals on the fixed-effects parameters. We could also use the parameter estimates and standard errors.
- The "estimates" (actually BLUPs) of the random effects can be considered as penalized estimates of these parameters in that they are shrunk towards the origin.
- Most of the prediction intervals for the random effects overlap zero.

Problems

3.1. Check the structure of documentation, structure and a summary of the `Orthodont` data set.

1. Create an `xyplot` of the `distance` versus `age` by `Subject` for the female subjects only. You can use the optional argument `subset = Sex == "Female"`

in the call to `xyplot` to achieve this. Use the optional argument `type = c("g","p","r")` to add reference lines to each panel.
2. Enhance the plot by choosing an aspect ratio for which the typical slope of the reference line is around 45^o. You can set it manually (something like `aspect = 4`) or with an automatic specification (`aspect = "xy"`). Change the layout so the panels form one row (`layout = c(11,1)`).
3. Order the panels according to increasing response at age 8. This is achieved with the optional argument `index.cond` which is a function of arguments `x` and `y`. In this case you could use `index.cond = function(x,y) y[x == 8]`. Add meaningful axis labels. Your final plot should be like



4. Fit a linear mixed model to the data for the females only with random effects for the intercept and for the slope by subject, allowing for correlation of these random effects within subject. Relate the fixed effects and the random effects' variances and covariances to the variability shown in the figure.
5. Produce a "caterpillar plot" of the random effects for intercept and slope. Does the plot indicate correlated random effects?
6. Consider what the Intercept coefficient and random effects represents. What will happen if you center the ages by subtracting 8 (the baseline year) or 11 (the middle of the age range)?
7. Repeat for the data from the male subjects.

3.2.
Fit a model to both the female and the male subjects in the `Orthodont` allowing for differences by sex in the fixed-effects for intercept (probably with respect to the centered age range) and slope.

Chapter 4
Computational methods

In this chapter we describe some of the details the computational methods for fitting mixed-effects models, as implemented in the `lme4` package, and the theoretical development behind these methods. We also provide the basis for later generalizations to models for non-Gaussian responses and to models in which the relationship between the conditional mean, μ, and the linear predictor, $\mathbf{X}\beta + \mathbf{Zb}$ is a nonlinear relationship.

This material is directed at those readers who wish to follow the theory and methodology of linear mixed models and how both can be extended to other forms of mixed models. Readers who are less interested in the "how" and the "why" of fitting mixed models than in the results themselves should not feel obligated to master these details.

4.1 Methods for linear mixed models

We begin by reviewing the definition of linear mixed-effects models and some of the basics of the computational methods, as given in Sect. 1.1.

4.1.1 Definitions and basic results.

As described in Sect. 1.1, a linear mixed-effects model is based on two vector-valued random variables: the q-dimensional vector of random effects, $\mathscr{B}$, and the n-dimensional response vector, $\mathscr{Y}$. The unconditional distribution of $\mathscr{B}$ and the conditional distribution of $\mathscr{Y}$, given $\mathscr{B} = \mathbf{b}$, are multivariate Gaussian distributions of the form

$$\begin{aligned} (\mathscr{Y}|\mathscr{B} = \mathbf{b}) &\sim \mathscr{N}(\mathbf{X}\beta + \mathbf{Zb}, \sigma^2\mathbf{I}) \\ \mathscr{B} &\sim \mathscr{N}(\mathbf{0}, \Sigma_\theta). \end{aligned} \tag{4.1}$$

The $q \times q$, symmetric, variance-covariance matrix, $\mathrm{Var}(\mathscr{B}) = \Sigma_\theta$, depends on the *variance-component parameter vector*, θ, and is *positive semidefinite*, which means that

$$\mathbf{b}^\mathsf{T}\Sigma_\theta\mathbf{b} \geq 0, \quad \forall \mathbf{b} \neq \mathbf{0}. \tag{4.2}$$

The symbol $\forall$ is read "for all". The fact that Σ_θ is positive semidefinite does not guarantee that Σ_θ^{-1} exists. We would need a stronger property, $\mathbf{b}^\mathsf{T}\Sigma_\theta\mathbf{b} > 0, \quad \forall \mathbf{b} \neq \mathbf{0}$, called positive definiteness, to ensure that Σ_θ^{-1} exists.

Many computational formulas for linear mixed models are written in terms of Σ_θ^{-1}. Such formulas will become unstable as Σ_θ approaches singularity, as it can. It is a fact that singular (i.e. non-invertible) Σ_θ can and do occur in practice, as we have seen in some of the examples in earlier chapters. Moreover, during the course of the numerical optimization by which the parameter estimates are determined, it is frequently the case that the deviance or the REML criterion will need to be evaluated at values of θ that produce a singular Σ_θ. Because of this we will take care to use computational methods that can be applied even when Σ_θ is singular and are stable as Σ_θ approaches singularity.

A relative covariance factor, Λ_θ, is any matrix that satisfies

$$\Sigma_\theta = \sigma^2\Lambda_\theta\Lambda_\theta^\mathsf{T}. \tag{4.3}$$

According to this definition, Σ depends on both σ and θ and we should write it as $\Sigma_{\sigma,\theta}$. However, we will blur that distinction and continue to write $\mathrm{Var}(\mathscr{B}) = \Sigma_\theta$. Another technicality is that the *common scale parameter*, σ, can, in theory, be zero. We will show that in practice the only way for its estimate, $\widehat{\sigma}$, to be zero is for the fitted values from the fixed-effects only, $\mathbf{X}\widehat{\beta}$, to be exactly equal to the observed data. This occurs only with data that have been (incorrectly) simulated without error. In practice we can safely assume that $\sigma > 0$. However, Λ_θ, like Σ_θ, can be singular.

Our computational methods are based on Λ_θ and do not require evaluation of Σ_θ. In fact, Σ_θ is explicitly evaluated only at the converged parameter estimates.

The spherical random effects, $\mathscr{U} \sim \mathscr{N}(\mathbf{0}, \sigma^2\mathbf{I}_q)$, determine $\mathscr{B}$ as

$$\mathscr{B} = \Lambda_\theta\mathscr{U}. \tag{4.4}$$

Although it may seem more reasonable to write $\mathscr{U}$ as a linear transformation of $\mathscr{B}$, we cannot do that when Λ_θ is singular, which is why (4.4) is in the form shown.

We can easily verify that (4.4) provides the desired distribution for $\mathscr{B}$. As a linear transformation of a multivariate Gaussian random variable, $\mathscr{B}$ will also be multivariate Gaussian. Its mean and variance-covariance matrix are straightforward to evaluate,

$$\begin{aligned}
\mathrm{E}[\mathscr{B}] &= \Lambda_\theta \mathrm{E}[\mathscr{U}] = \Lambda_\theta \mathbf{0} = \mathbf{0} && (4.5)\\
\mathrm{Var}(\mathscr{B}) &= \mathrm{E}\left[(\mathscr{B} - \mathrm{E}[\mathscr{B}])(\mathscr{B} - \mathrm{E}[\mathscr{B}])^\mathsf{T}\right] = \mathrm{E}\left[\mathscr{B}\mathscr{B}^\mathsf{T}\right] && (4.6)\\
&= \mathrm{E}\left[\Lambda_\theta \mathscr{U}\mathscr{U}^\mathsf{T}\Lambda_\theta^\mathsf{T}\right] = \Lambda_\theta \mathrm{E}[\mathscr{U}\mathscr{U}^\mathsf{T}]\Lambda_\theta^\mathsf{T} = \Lambda_\theta \mathrm{Var}(\mathscr{U})\Lambda_\theta^\mathsf{T}\\
&= \Lambda_\theta \sigma^2 \mathbf{I}_q \Lambda_\theta^\mathsf{T} = \sigma^2 \Lambda_\theta \Lambda_\theta^\mathsf{T} = \Sigma_\theta
\end{aligned}$$

and have the desired form.

Just as we concentrate on how θ determines Λ_θ, not Σ_θ, we will concentrate on properties of $\mathscr{U}$ rather than $\mathscr{B}$. In particular, we now define the model according to the distributions

$$\begin{aligned}
(\mathscr{Y}|\mathscr{U} = \mathbf{u}) &\sim \mathscr{N}(\mathbf{Z}\Lambda_\theta \mathbf{u} + \mathbf{X}\beta, \sigma^2 \mathbf{I}_n)\\
\mathscr{U} &\sim \mathscr{N}(\mathbf{0}, \sigma^2 \mathbf{I}_q).
\end{aligned} \qquad (4.7)$$

To allow for extensions to other types of mixed models we distinguish between the *linear predictor*

$$\gamma = \mathbf{Z}\Lambda_\theta \mathbf{u} + \mathbf{X}\beta \qquad (4.8)$$

and the *conditional mean* of $\mathscr{Y}$, given $\mathscr{U} = \mathbf{u}$, which is

$$\mu = \mathrm{E}[\mathscr{Y}|\mathscr{U} = \mathbf{u}]. \qquad (4.9)$$

For a linear mixed model $\mu = \gamma$. In other forms of mixed models the conditional mean, μ, can be a nonlinear function of the linear predictor, γ. For some models the dimension of γ is a multiple of n, the dimension of μ and $\mathbf{y}$, but for a linear mixed model the dimension of γ must be n. Hence, the model matrices $\mathbf{Z}$ must be $n \times q$ and $\mathbf{X}$ must be $n \times p$.

4.1.2 The conditional distribution $(\mathscr{U}|\mathscr{Y} = \mathbf{y})$

In this section it will help to be able to distinguish between the observed response vector and an arbitrary value of $\mathscr{Y}$. For this section only we will write the observed data vector as $\mathbf{y}_{\mathrm{obs}}$, whereas $\mathbf{y}$ without the subscript will refer to an arbitrary value of the random variable $\mathscr{Y}$.

The likelihood of the parameters, θ, β, and σ, given the observed data, $\mathbf{y}_{\mathrm{obs}}$, is the probability density of $\mathscr{Y}$, evaluated at $\mathbf{y}_{\mathrm{obs}}$. The numerical values of the probability density and the likelihood are the same but the interpretation is different. For the density we think of the parameters as being fixed and the value of $\mathbf{y}$ as the variable. For the likelihood we think of $\mathbf{y}$ as fixed at the observed vector of responses and the parameters, θ, β and σ as varying.

The natural approach for evaluating the likelihood is to determine the marginal distribution of $\mathscr{Y}$, which in this case amounts to determining the

marginal density of $\mathscr{Y}$, and evaluate that density at $\mathbf{y}_{\text{obs}}$. To follow this course we would first determine the joint density of $\mathscr{U}$ and $\mathscr{Y}$, written $f_{\mathscr{U},\mathscr{Y}}(\mathbf{u},\mathbf{y})$, then integrate this density with respect $\mathbf{u}$ to create the marginal density, $f_{\mathscr{Y}}(\mathbf{y})$, and finally evaluate this marginal density at $\mathbf{y}_{\text{obs}}$.

To allow for later generalizations we will change the order of these steps slightly. We evaluate the joint density function, $f_{\mathscr{U},\mathscr{Y}}(\mathbf{u},\mathbf{y})$, at $\mathbf{y}_{\text{obs}}$, producing the *unnormalized conditional density*, $h(\mathbf{u})$. We say that h is "unnormalized" because the conditional density is a multiple of h

$$f_{\mathscr{U}|\mathscr{Y}}(\mathbf{u}|\mathbf{y}_{\text{obs}}) = \frac{h(\mathbf{u})}{\int_{\mathbb{R}^q} h(\mathbf{u})\,d\mathbf{u}}. \tag{4.10}$$

In some theoretical developments the normalizing constant, which is the integral in the denominator of an expression like (4.10), is not of interest. Here it is of interest because the normalizing constant is exactly the likelihood that we wish to evaluate,

$$L(\theta,\beta,\sigma|\mathbf{y}_{\text{obs}}) = \int_{\mathbb{R}^q} h(\mathbf{u})\,d\mathbf{u}. \tag{4.11}$$

For a linear mixed model, where all the distributions of interest are multivariate Gaussian and the conditional mean, μ, is a linear function of both $\mathbf{u}$ and β, the distinction between evaluating the joint density at $\mathbf{y}_{\text{obs}}$ to produce $h(\mathbf{u})$ then integrating with respect to $\mathbf{u}$, as opposed to first integrating the joint density then evaluating at $\mathbf{y}_{\text{obs}}$ is not terribly important. For other mixed models this distinction can be important. In particular, generalized linear mixed models, described in Sect. **??**, are often used to model a discrete response, such as a binary response or a count, leading to a joint distribution for $\mathscr{Y}$ and $\mathscr{U}$ that is discrete with respect to one variable, $\mathbf{y}$, and continuous with respect to the other, $\mathbf{u}$. In such cases there isn't a joint density for $\mathscr{Y}$ and $\mathscr{U}$. The necessary distribution theory for general $\mathbf{y}$ and $\mathbf{u}$ is well-defined but somewhat awkward to describe. It is much easier to realize that we are only interested in the observed response vector, $\mathbf{y}_{\text{obs}}$, not some arbitrary value of $\mathbf{y}$, so we can concentrate on the conditional distribution of $\mathscr{U}$ given $\mathscr{Y} = \mathbf{y}_{\text{obs}}$. For all the mixed models we will consider, the conditional distribution, $(\mathscr{U}|\mathscr{Y} = \mathbf{y}_{\text{obs}})$, is continuous and both the conditional density, $f_{\mathscr{U}|\mathscr{Y}}(\mathbf{u}|\mathbf{y}_{\text{obs}})$ and its unnormalized form, $h(\mathbf{u})$, are well-defined.

4.1.3 Integrating $h(\mathbf{u})$ in the linear mixed model

The integral defining the likelihood in (4.11) has a closed form in the case of a linear mixed model but not for some of the more general forms of mixed models. To motivate methods for approximating the likelihood in more general situations, we describe in some detail how the integral can be evaluated

using the sparse Cholesky factor, $\mathbf{L}_\theta$, and the conditional mode,

$$\tilde{\mathbf{u}} = \arg\max_{\mathbf{u}} f_{\mathscr{U}|\mathscr{Y}}(\mathbf{u}|\mathbf{y}) = \arg\max_{\mathbf{u}} h(\mathbf{u}) = \arg\max_{\mathbf{u}} f_{\mathscr{Y}|\mathscr{U}}(\mathbf{y}|\mathbf{u})\, f_{\mathscr{U}}(\mathbf{u}) \tag{4.12}$$

The notation $\arg\max_{\mathbf{u}}$ means that $\tilde{\mathbf{u}}$ is the value of $\mathbf{u}$ that maximizes the expression on the right.

In general, the *mode* of a continuous distribution is the value that maximizes the density. The value $\tilde{\mathbf{u}}$ is called the conditional mode of $\mathbf{u}$, given the observed $\mathbf{y}$, because it maximizes the conditional density of $\mathscr{U}$ given $\mathscr{Y} = \mathbf{y}$. The location of the maximum can also be determined by maximizing the unnormalized conditional density because $h(\mathbf{u})$ is just a constant multiple of $f_{\mathscr{U}|\mathscr{Y}}(\mathbf{y}|\mathbf{u})$. The last part of (4.12) is simply a re-expression of $h(\mathbf{u})$ as the product of $f_{\mathscr{Y}|\mathscr{U}}(\mathbf{y}|\mathbf{u})$ and $f_{\mathscr{U}}(\mathbf{u})$. For a linear mixed model these densities are

$$f_{\mathscr{Y}|\mathscr{U}}(\mathbf{y}|\mathbf{u}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\|\mathbf{y}-\mathbf{X}\beta-\mathbf{Z}\Lambda_\theta\mathbf{u}\|^2}{2\sigma^2}\right) \tag{4.13}$$

$$f_{\mathscr{U}}(\mathbf{u}) = \frac{1}{(2\pi\sigma^2)^{q/2}} \exp\left(-\frac{\|\mathbf{u}\|^2}{2\sigma^2}\right) \tag{4.14}$$

It is easiest to consider the product of these densities on the deviance scale (negative twice the logarithm of the density).

$$-2\log\left(h(\mathbf{u})\right) = (n+q)\log(2\pi\sigma^2) + \frac{\|\mathbf{y}-\mathbf{X}\beta-\mathbf{Z}\Lambda_\theta\mathbf{u}\|^2 + \|\mathbf{u}\|^2}{\sigma^2}. \tag{4.15}$$

Because of the negative sign, $\tilde{\mathbf{u}}$ will be the value of $\mathbf{u}$ that minimizes the expression on the right of (4.15).

The only part of the expression in (4.15) that depends on $\mathbf{u}$ is the numerator of the second term. Thus

$$\tilde{\mathbf{u}} = \arg\min_{\mathbf{u}} \|\mathbf{y}-\mathbf{X}\beta-\mathbf{Z}\Lambda_\theta\mathbf{u}\|^2 + \|\mathbf{u}\|^2. \tag{4.16}$$

The expression to be minimized, called the *objective function*, is described as a *penalized residual sum of squares* (PRSS) and the minimizer, $\tilde{\mathbf{u}}$ is called the *penalized least squares* (PLS) solution. They are given these names because the first term in the objective, $\|\mathbf{y}-\mathbf{X}\beta-\mathbf{Z}\Lambda_\theta\mathbf{u}\|^2$, is a sum of squared residuals, and the second term, $\|\mathbf{u}\|^2$, is a penalty on the length, $\|\mathbf{u}\|$, of $\mathbf{u}$. Larger values of $\mathbf{u}$ (in the sense of greater lengths as vectors) incur a higher penalty.

The PRSS criterion determining the conditional mode balances fidelity to the observed data, producing a small residual sum of squares, against simplicity of the model, small values of $\|\mathbf{u}\|$. In this sense it is the type of criterion called a smoothing objective.

For the purpose of evaluating the likelihood we will regard the PRSS criterion as a function of the parameters, given the data, and write its minimum

value as

$$r^2_{\theta,\beta} = \min_{\mathbf{u}} \|\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\Lambda_\theta \mathbf{u}\|^2 + \|\mathbf{u}\|^2. \tag{4.17}$$

Notice that β is only involved in the linear predictor expression in (4.17). We will see that $\tilde{\mathbf{u}}$ can be determined by a direct (i.e. non-iterative) calculation and, in fact, we can minimize the PRSS criterion with respect to $\mathbf{u}$ and β simultaneously without iterating. We write this minimum value as

$$r^2_{\theta} = \min_{\mathbf{u},\beta} \|\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\Lambda_\theta \mathbf{u}\|^2 + \|\mathbf{u}\|^2. \tag{4.18}$$

The value of β at the minimum is called the conditional estimate of β given θ, written $\widehat{\beta}_\theta$.

4.1.4 Determining the PLS Solutions, $\tilde{\mathbf{u}}$ and $\widehat{\beta}_\theta$

One way of expressing a penalized least squares problem like (4.17) is by incorporating the penalty as "pseudo-data" in a standard least squares problem. We extend the "response vector", which is $\mathbf{y} - \mathbf{X}\beta$ when we minimize with respect to $\mathbf{u}$ only, with q responses that are 0 and we extend the predictor expression, $\mathbf{Z}\Lambda_\theta\mathbf{u}$ with $\mathbf{I}_q\mathbf{u}$. Writing this as a least squares problem produces

$$\tilde{\mathbf{u}} = \arg\min_{\mathbf{u}} \left\| \begin{bmatrix} \mathbf{y} - \mathbf{X}\beta \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{Z}\Lambda_\theta \\ \mathbf{I}_q \end{bmatrix} \mathbf{u} \right\|^2 \tag{4.19}$$

with a solution that satisfies

$$\left(\Lambda_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{Z}\Lambda_\theta + \mathbf{I}_q\right)\tilde{\mathbf{u}} = \Lambda_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\left[\mathbf{y} - \mathbf{X}\beta\right] \tag{4.20}$$

To evaluate $\tilde{\mathbf{u}}$ we form the *sparse Cholesky factor*, $\mathbf{L}_\theta$, which is a lower triangular $q \times q$ matrix that satisfies

$$\mathbf{L}_\theta\mathbf{L}_\theta^\mathsf{T} = \Lambda_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{Z}\Lambda_\theta + \mathbf{I}_q. \tag{4.21}$$

In earlier chapters we have seen the often the random effects vector is re-ordered before $\mathbf{L}_\theta$ is created. The re-ordering or permutation of the elements of $\mathbf{u}$ and, correspondingly the columns of the model matrix, $\mathbf{Z}\blacksquare_\theta$, does not affect the theory of linear mixed models but can have a profound effect on the time and storage required to evaluate $\mathbf{L}_\theta$ in large problems. We write the effect of the permutation as multiplication by a $q \times q$ *permutation matrix*, $\mathbf{P}$, although in practice we perform the permutation without ever constructing $\mathbf{P}$. The matrix is for notation only. It is determined by the structure of $\Lambda_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{Z}\Lambda_\theta + \mathbf{I}_q$ for the initial value of θ in the optimization of the profiled

deviance. The permutation itself does not depend on θ as long as θ is not on the boundary of the parameter space, where it generates a singular Λ_θ.

Thus we alter the definition of the sparse Cholesky factor to be the sparse lower triangular $q \times q$ matrix that satisfies

$$\mathbf{L}_\theta \mathbf{L}_\theta^\mathsf{T} = \mathbf{P}\Lambda_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{Z}\Lambda_\theta\mathbf{P}^\mathsf{T} + \mathbf{I}_q. \tag{4.22}$$

Many sparse matrix methods, including the sparse Cholesky decomposition, are performed in two stages: the *symbolic phase* in which the locations of the non-zeros in the result are determined and the *numeric phase* in which the numeric values at these positions are evaluated. The symbolic phase for the decomposition (4.22), which includes determining the permutation, $\mathbf{P}$, need only be done once. Evaluation of $\mathbf{L}_\theta$ for subsequent values of θ requires only the numeric phase, which is often much faster than the symbolic phase.

The permutation, $\mathbf{P}$, serves two purposes. The first and most important purpose is to reduce the number of non-zeros in the factor, $\mathbf{L}_\theta$. The factor is non-zero at every non-zero of the lower triangle of the matrix being decomposed. However, as we saw in Fig. 2.4 of Sect. 2.1.2, there may be positions in the factor that get filled-in even though they are known to be zero in the matrix being decomposed. The *fill-reducing permutation* is chosen according to certain heuristics to reduce the amount of fill-in. We use the approximate minimal degree (AMD) method described in **?**. After the fill-reducing permutation is determined, a "post-ordering" is applied. This has the effect of concentrating the non-zeros near the diagonal of the factor. See Davis [2006] for more details.

The transpose of a permutation matrix like $\mathbf{P}$ is its inverse. That is $\mathbf{P}^\mathsf{T}\mathbf{P} = \mathbf{P}\mathbf{P}^\mathsf{T} = \mathbf{I}_q$. Applying the permutation corresponds to multiplying a vector or matrix by $\mathbf{P}$. Applying the inverse permutation corresponds to multiplication by $\mathbf{P}^\mathsf{T}$.

Obtaining the Cholesky factor, $\mathbf{L}_\theta$, may not seem to be great progress toward determining $\tilde{\mathbf{u}}$ because we still must solve

$$\mathbf{L}_\theta \mathbf{L}_\theta^\mathsf{T} \tilde{\mathbf{u}} = \mathbf{P}\Lambda_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}[\mathbf{y} - \mathbf{X}\beta] \tag{4.23}$$

for $\tilde{\mathbf{u}}$. However, this is the key step in computational methods in the `lme4` package. The ability to evaluate $\mathbf{L}_\theta$ rapidly for many different values of θ is what makes the computational methods in `lme4` feasible, even when applied to very large data sets with complex structure. Determining $\tilde{\mathbf{u}}$ through (4.23) is a straightforward process because $\mathbf{L}_\theta$ is lower triangular.

After evaluating $\mathbf{L}_\theta$ and using that to solve for $\tilde{\mathbf{u}}$, which also produces $r^2_{\beta,\theta}$, we can write the PRSS for a general $\mathbf{u}$ as

$$\|\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\Lambda_\theta\mathbf{u}\|^2 + \|\mathbf{u}\|^2 = r^2_{\theta,\beta} + \|\mathbf{L}_\theta^\mathsf{T}(\mathbf{u} - \tilde{\mathbf{u}})\|^2 \tag{4.24}$$

which finally allows us to evaluate the likelihood. We plug the right hand side of (4.24) into the definition of $h(\mathbf{u})$ and apply a change of variable to

$$\mathbf{z} = \frac{\mathbf{L}_\theta^\top(\mathbf{u} - \tilde{\mathbf{u}})}{\sigma}. \tag{4.25}$$

The determinant of the Jacobian of this transformation,

$$\left|\frac{d\mathbf{z}}{d\mathbf{u}}\right| = \left|\frac{\mathbf{L}_\theta^\top}{\sigma}\right| = \frac{|\mathbf{L}_\theta|}{\sigma^q} \tag{4.26}$$

is required for the change of variable in the integral. We use the letter $\mathbf{z}$ for the transformed value because we will rearrange the integral to have the form of the integral of the density of the standard multivariate normal distribution.

Putting all this together gives

$$\begin{aligned}
L(\theta,\beta,\sigma) &= \int_{\mathbb{R}^q} h(\mathbf{u})\,d\mathbf{u} \\
&= \int_{\mathbb{R}^q} \frac{1}{(2\pi\sigma^2)^{(n+q)/2}} \exp\left(-\frac{r_{\theta,\beta}^2 + \|\mathbf{L}_\theta^\top(\mathbf{u}-\tilde{\mathbf{u}})\|^2}{2\sigma^2}\right) d\mathbf{u} \\
&= \frac{\exp\left(-\frac{r_{\theta,\beta}^2}{2\sigma^2}\right)}{(2\pi\sigma^2)^{n/2}} \int_{\mathbb{R}^q} \frac{1}{(2\pi)^{q/2}} \exp\left(-\frac{\|\mathbf{L}_\theta^\top(\mathbf{u}-\tilde{\mathbf{u}})\|^2}{2\sigma^2}\right) \frac{|\mathbf{L}_\theta|}{|\mathbf{L}_\theta|}\frac{d\mathbf{u}}{\sigma^q} \\
&= \frac{\exp\left(-\frac{r_{\theta,\beta}^2}{2\sigma^2}\right)}{(2\pi\sigma^2)^{n/2}|\mathbf{L}_\theta|} \int_{\mathbb{R}^q} \frac{e^{-\|\mathbf{z}\|^2/2}}{(2\pi)^{q/2}}\,d\mathbf{z} \\
&= \frac{\exp\left(-\frac{r_{\theta,\beta}^2}{2\sigma^2}\right)}{(2\pi\sigma^2)^{n/2}|\mathbf{L}_\theta|}.
\end{aligned} \tag{4.27}$$

The deviance, which is negative twice the (natural) logarithm of the likelihood, becomes

$$d(\theta,\beta,\sigma|\mathbf{y}) = n\log(2\pi\sigma^2) + \log(|\mathbf{L}_\theta|^2) + \frac{r_{\beta,\theta}^2}{\sigma^2}. \tag{4.28}$$

The maximum likelihood estimates of the parameters are those that minimize the deviance, (4.28).

Evaluating the determinant, $|\mathbf{L}_\theta|$, may seem formidable but, because $\mathbf{L}$ is triangular, its determinant is simply the product of its diagonal elements.

Equation (4.28) is a remarkably compact expression, considering that the class of models to which it applies is very large indeed. However, we can do better than this if we notice that β enters (4.28) only through the PRSS $r_{\beta,\theta}^2$, and, for any value of θ, minimizing this expression with respect to β is just another least squares problem. Let $\widehat{\beta}_\theta$ be the value of β that minimizes the PRSS simultaneously with respect to β and $\mathbf{u}$ and let r_θ^2 be the minimum

PRSS. Furthermore, if we set $\widehat{\sigma^2}_\theta = r_\theta^2/n$, which is the value of σ^2 that minimizes the deviance for a given value of r_θ^2, then the *profiled deviance*, which is a function of θ only, is

$$\tilde{d}(\theta|\mathbf{y}) = \log(|\mathbf{L}_\theta|^2) + n\left[1 + \log\left(\frac{2\pi r_\theta^2}{n}\right)\right]. \tag{4.29}$$

The profiled deviance (4.29) is optimized numerically with respect to θ to determine the MLE, $\widehat{\theta}$. The MLEs for the other parameters, $\widehat{\beta}$ and $\widehat{\sigma}$ are the conditional estimates evaluated at $\widehat{\theta}$.

4.2 Generalizations to Other Mixed Models

The `lme4` package provides `R` functions to fit and analyze linear mixed models, generalized linear mixed models and nonlinear mixed models. These models are called *mixed-effects models* or, more simply, *mixed models* because they incorporate both *fixed-effects* parameters, which apply to an entire population or to certain well-defined and repeatable subsets of a population, and *random effects*, which apply to the particular experimental units or observational units in the study. Such models are also called *multilevel* models because the random effects represent levels of variation in addition to the per-observation noise term that is incorporated in common statistical models such as linear regression models, generalized linear models and nonlinear regression models.

We begin by describing common properties of these mixed models and the general computational approach used in the `lme4` package. The estimates of the parameters in a mixed model are determined as the values that optimize an objective function — either the likelihood of the parameters given the observed data, for maximum likelihood (ML) estimates, or a related objective function called the REML criterion. Because this objective function must be evaluated at many different values of the model parameters during the optimization process, we focus on the evaluation of the objective function and a critical computation in this evalution — determining the solution to a penalized, weighted least squares (PWLS) problem.

The dimension of the solution of the PWLS problem can be very large, perhaps in the millions. Furthermore, such problems must be solved repeatedly during the optimization process to determine parameter estimates. The whole approach would be infeasible were it not for the fact that the matrices determining the PWLS problem are sparse and we can use sparse matrix storage formats and sparse matrix computations [Davis, 2006]. In particular, the whole computational approach hinges on the extraordinarily efficient methods for determining the Cholesky decomposition of sparse, symmetric, positive-definite matrices embodied in the CHOLMOD library of C functions [Davis, 2005].

In the next section we describe the general form of the mixed models that can be represented in the `lme4` package and the computational approach embodied in the package. In the following section we describe a particular form of mixed model, called a linear mixed model, and the computational details for those models. In the fourth section we describe computational methods for generalized linear mixed models, nonlinear mixed models and generalized nonlinear mixed models.

4.3 Formulation of mixed models

A mixed-effects model incorporates two vector-valued random variables: the n-dimensional response vector, $\mathscr{Y}$, and the q-dimensional random effects vector, $\mathscr{B}$. We observe the value, $\mathbf{y}$, of $\mathscr{Y}$. We do not observe the value of $\mathscr{B}$.

The random variable $\mathscr{Y}$ may be continuous or discrete. That is, the observed data, $\mathbf{y}$, may be on a continuous scale or they may be on a discrete scale, such as binary responses or responses representing a count. In our formulation, the random variable $\mathscr{B}$ is always continous.

We specify a mixed model by describing the unconditional distribution of $\mathscr{B}$ and the conditional distribution $(\mathscr{Y}|\mathscr{B}=\mathbf{b})$.

4.3.1 The unconditional distribution of $\mathscr{B}$

In our formulation, the unconditional distribution of $\mathscr{B}$ is always a q-dimensional multivariate Gaussian (or "normal") distribution with mean $\mathbf{0}$ and with a parameterized covariance matrix,

$$\mathscr{B} \sim \mathscr{N}\left(\mathbf{0}, \sigma^2 \Lambda(\theta)\Lambda^{\mathsf{T}}(\theta)\right). \tag{4.30}$$

The scalar, σ, in (4.30), is called the *common scale parameter*. As we will see later, not all types of mixed models incorporate this parameter. We will include σ^2 in the general form of the unconditional distribution of $\mathscr{B}$ with the understanding that, in some models, $\sigma \equiv 1$.

The $q \times q$ matrix $\Lambda(\theta)$, which is a left factor of the covariance matrix (when $\sigma = 1$) or the relative covariance matrix (when $\sigma \neq 1$), depends on an m-dimensional parameter θ. Typically $m \ll q$; in the examples we show below it is always the case that $m < 5$, even when q is in the thousands. The fact that m is very small is important because, as we shall see, determining the parameter estimates in a mixed model can be expressed as an optimization problem with respect to θ only.

The parameter θ may be, and typically is, subject to constraints. For ease of computation, we require that the constraints be expressed as "box"

constraints of the form $\theta_{iL} \le \theta_i \le \theta_{iU}, i = 1,\ldots,m$ for constants θ_{iL} and $\theta_{iU}, i = 1,\ldots,m$. We shall write the set of such constraints as $\theta_L \le \theta \le \theta_R$. The matrix $\Lambda(\theta)$ is required to be non-singular (i.e. invertible) when θ is not on the boundary.

4.3.2 The conditional distribution, $(\mathscr{Y}|\mathscr{B} = \mathbf{b})$

The conditional distribution, $(\mathscr{Y}|\mathscr{B} = \mathbf{b})$, must satisfy:

1. The conditional mean, $\mu_{\mathscr{Y}|\mathscr{B}}(\mathbf{b}) = \mathrm{E}[\mathscr{Y}|\mathscr{B} = \mathbf{b}]$, depends on $\mathbf{b}$ only through the value of the *linear predictor*, $\mathbf{Zb} + \mathbf{X}\beta$, where β is the p-dimensional *fixed-effects* parameter vector and the *model matrices*, $\mathbf{Z}$ and $\mathbf{X}$, are fixed matrices of the appropriate dimension. That is, the two model matrices must have the same number of rows and must have q and p columns, respectively. The number of rows in $\mathbf{Z}$ and $\mathbf{X}$ is a multiple of n, the dimension of $\mathbf{y}$.
2. The scalar distributions, $(\mathscr{Y}_i|\mathscr{B} = \mathbf{b}), i = 1,\ldots,n$, all have the same form and are completely determined by the conditional mean, $\mu_{\mathscr{Y}|\mathscr{B}}(\mathbf{b})$ and, at most, one additional parameter, σ, which is the common scale parameter.
3. The scalar distributions, $(\mathscr{Y}_i|\mathscr{B} = \mathbf{b}), i = 1,\ldots,n$, are independent. That is, the components of $\mathscr{Y}$ are *conditionally independent* given $\mathscr{B}$.

An important special case of the conditional distribution is the multivariate Gaussian distribution of the form

$$(\mathscr{Y}|\mathscr{B} = \mathbf{b}) \sim \mathscr{N}(\mathbf{Zb} + \mathbf{X}\beta, \sigma^2\mathbf{I}_n) \tag{4.31}$$

where $\mathbf{I}_n$ denotes the identity matrix of size n. In this case the conditional mean, $\mu_{\mathscr{Y}|\mathscr{B}}(\mathbf{b})$, is exactly the linear predictor, $\mathbf{Zb} + \mathbf{X}\beta$, a situation we will later describe as being an "identity link" between the conditional mean and the linear predictor. Models with conditional distribution (4.31) are called *linear mixed models*.

4.3.3 A change of variable to "spherical" random effects

Because the conditional distribution $(\mathscr{Y}|\mathscr{B} = \mathbf{b})$ depends on $\mathbf{b}$ only through the linear predictor, it is easy to express the model in terms of a linear transformation of $\mathscr{B}$. We define the linear transformation from a q-dimensional "spherical" Gaussian random variable, $\mathscr{U}$, to $\mathscr{B}$ as

$$\mathscr{B} = \Lambda(\theta)\mathscr{U}, \quad \mathscr{U} \sim \mathscr{N}(\mathbf{0}, \sigma^2\mathbf{I}_q). \tag{4.32}$$

(The term "spherical" refers to the fact that contours of constant probability density for $\mathscr{U}$ are spheres centered at the mean — in this case, $\mathbf{0}$.)

When θ is not on the boundary this is an invertible transformation. When θ is on the boundary the transformation can fail to be invertible. However, we will only need to be able to express $\mathscr{B}$ in terms of $\mathscr{U}$ and that transformation is well-defined, even when θ is on the boundary.

The linear predictor, as a function of $\mathbf{u}$, is

$$\eta(\mathbf{u}) = \mathbf{Z}\Lambda(\theta)\mathbf{u} + \mathbf{X}\beta. \tag{4.33}$$

When we wish to emphasize the role of the model parameters, θ and β, in the formulation of η, we will write the linear predictor as $\eta(\mathbf{u}, \theta, \beta)$.

4.3.4 The conditional density $(\mathscr{U}|\mathscr{Y} = \mathbf{y})$

Because we observe $\mathbf{y}$ and do not observe $\mathbf{b}$ or $\mathbf{u}$, the conditional distribution of interest, for the purposes of statistical inference, is $(\mathscr{U}|\mathscr{Y} = \mathbf{y})$ (or, equivalently, $(\mathscr{B}|\mathscr{Y} = \mathbf{y})$). This conditional distribution is always a continuous distribution with conditional probability density $f_{\mathscr{U}|\mathscr{Y}}(\mathbf{u}|\mathbf{y})$.

We can evaluate $f_{\mathscr{U}|\mathscr{Y}}(\mathbf{u}|\mathbf{y})$, up to a constant, as the product of the unconditional density, $f_{\mathscr{U}}(\mathbf{u})$, and the conditional density (or the probability mass function, whichever is appropriate), $f_{\mathscr{Y}|\mathscr{U}}(\mathbf{y}|\mathbf{u})$. We write this unnormalized conditional density as

$$h(\mathbf{u}|\mathbf{y}, \theta, \beta, \sigma) = f_{\mathscr{Y}|\mathscr{U}}(\mathbf{y}|\mathbf{u}, \theta, \beta, \sigma) f_{\mathscr{U}}(\mathbf{u}|\sigma). \tag{4.34}$$

We say that h is the "unnormalized" conditional density because all we know is that the conditional density is proportional to $h(\mathbf{u}|\mathbf{y}, \theta, \beta, \sigma)$. To obtain the conditional density we must normalize h by dividing by the value of the integral

$$L(\theta, \beta, \sigma|\mathbf{y}) = \int_{\mathbb{R}^q} h(\mathbf{u}|\mathbf{y}, \theta, \beta, \sigma)\, d\mathbf{u}. \tag{4.35}$$

We write the value of the integral (4.35) as $L(\theta, \beta, \sigma|\mathbf{y})$ because it is exactly the *likelihood* of the parameters θ, β and σ, given the observed data $\mathbf{y}$. The *maximum likelihood (ML) estimates* of these parameters are the values that maximize L.

4.3.5 Determining the ML estimates

The general problem of maximizing $L(\theta, \beta, \sigma|\mathbf{y})$ with respect to θ, β and σ can be formidable because each evaluation of this function involves a po-

tentially high-dimensional integral and because the dimension of $\boldsymbol{\beta}$ can be large. However, this general optimization problem can be split into manageable subproblems. Given a value of $\boldsymbol{\theta}$ we can determine the *conditional mode*, $\tilde{\mathbf{u}}(\boldsymbol{\theta})$, of $\mathbf{u}$ and the *conditional estimate*, $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ simultaneously using *penalized, iteratively re-weighted least squares* (PIRLS). The conditional mode and the conditional estimate are defined as

$$\begin{bmatrix}\tilde{\mathbf{u}}(\boldsymbol{\theta})\\ \tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})\end{bmatrix} = \arg\max_{\mathbf{u},\boldsymbol{\beta}} h(\mathbf{u}|\mathbf{y},\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\sigma}). \tag{4.36}$$

(It may look as if we have missed the dependence on $\boldsymbol{\sigma}$ on the left-hand side but it turns out that the scale parameter does not affect the location of the optimal values of quantities in the linear predictor.)

As is common in such optimization problems, we re-express the conditional density on the *deviance scale*, which is negative twice the logarithm of the density, where the optimization becomes

$$\begin{bmatrix}\tilde{\mathbf{u}}(\boldsymbol{\theta})\\ \tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})\end{bmatrix} = \arg\min_{\mathbf{u},\boldsymbol{\beta}} -2\log\left(h(\mathbf{u}|\mathbf{y},\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\sigma})\right). \tag{4.37}$$

It is this optimization problem that can be solved quite efficiently using PIRLS. In fact, for linear mixed models, which are described in the next section, $\tilde{\mathbf{u}}(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ can be directly evaluated.

The second-order Taylor series expansion of $-2\log h$ at $\tilde{\mathbf{u}}(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ provides the Laplace approximation to the profiled deviance. Optimizing this function with respect to $\boldsymbol{\theta}$ provides the ML estimates of $\boldsymbol{\theta}$, from which the ML estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$ (if used) are derived.

4.4 Methods for linear mixed models

As indicated in the introduction, a critical step in our methods for determining the maximum likelihood estimates of the parameters in a mixed model is solving a penalized, weighted least squares (PWLS) problem. We will motivate the general form of the PWLS problem by first considering computational methods for linear mixed models that result in a penalized least squares (PLS) problem.

Recall from §4.3.2 that, in a linear mixed model, both the conditional distribution, $(\mathscr{Y}|\mathscr{U}=\mathbf{u})$, and the unconditional distribution, $\mathscr{U}$, are spherical Gaussian distributions and that the conditional mean, $\boldsymbol{\mu}_{\mathscr{Y}|\mathscr{U}}(\mathbf{u})$, is the linear predictor, $\boldsymbol{\gamma}(\mathbf{u})$. Because all the distributions determining the model are continuous distributions, we consider their densities. On the deviance scale these are

$$\begin{aligned}
-2\log(f_{\mathscr{U}}(\mathbf{u})) &= q\log(2\pi\sigma^2) + \frac{\|\mathbf{u}\|^2}{\sigma^2} \\
-2\log(f_{\mathscr{Y}|\mathscr{U}}(\mathbf{y}|\mathbf{u})) &= n\log(2\pi\sigma^2) + \frac{\|\mathbf{y}-\mathbf{Z}\Lambda(\theta)\mathbf{u}-\mathbf{X}\beta\|^2}{\sigma^2} \\
-2\log(h(\mathbf{u}|\mathbf{y},\theta,\beta,\sigma)) &= (n+q)\log(2\pi\sigma^2) + \frac{\|\mathbf{y}-\gamma(\mathbf{u},\theta,\beta)\|^2+\|\mathbf{u}\|^2}{\sigma^2} \\
&= (n+q)\log(2\pi\sigma^2) + \frac{d(\mathbf{u}|\mathbf{y},\theta,\beta)}{\sigma^2}
\end{aligned} \tag{4.38}$$

In (4.38) the *discrepancy* function,

$$d(\mathbf{u}|\mathbf{y},\theta,\beta) = \|\mathbf{y}-\gamma(\mathbf{u},\theta,\beta)\|^2 + \|\mathbf{u}\|^2 \tag{4.39}$$

has the form of a penalized residual sum of squares in that the first term, $\|\mathbf{y}-\gamma(\mathbf{u},\theta,\beta)\|^2$ is the residual sum of squares for $\mathbf{y}$, $\mathbf{u}$, θ and β and the second term, $\|\mathbf{u}\|^2$, is a penalty on the size of $\mathbf{u}$. Notice that the discrepancy does not depend on the common scale parameter, σ.

4.4.1 The canonical form of the discrepancy

Using a so-called "pseudo data" representation, we can write the discrepancy as a residual sum of squares for a regression model that is linear in both $\mathbf{u}$ and β

$$d(\mathbf{u}|\mathbf{y},\theta,\beta) = \left\| \begin{bmatrix}\mathbf{y}\\ \mathbf{0}\end{bmatrix} - \begin{bmatrix}\mathbf{Z}\Lambda(\theta) & \mathbf{X}\\ \mathbf{I}_q & \mathbf{0}\end{bmatrix} \begin{bmatrix}\mathbf{u}\\ \beta\end{bmatrix} \right\|^2. \tag{4.40}$$

The term "pseudo data" reflects the fact that we have added q "pseudo observations" to the observed response, $\mathbf{y}$, and to the linear predictor, $\gamma(\mathbf{u},\theta,\beta) = \mathbf{Z}\Lambda(\theta)\mathbf{u}+\mathbf{X}\beta$, in such a way that their contribution to the overall residual sum of squares is exactly the penalty term in the discrepancy.

In the form (4.40) we can see that the discrepancy is a quadratic form in both $\mathbf{u}$ and β. Furthermore, because we require that $\mathbf{X}$ has full column rank, the discrepancy is a positive-definite quadratic form in $\mathbf{u}$ and β that is minimized at $\tilde{\mathbf{u}}(\theta)$ and $\tilde{\beta}(\theta)$ satisfying

$$\begin{bmatrix}\Lambda^\mathsf{T}(\theta)\mathbf{Z}^\mathsf{T}\mathbf{Z}\Lambda(\theta)+\mathbf{I}_q & \Lambda^\mathsf{T}(\theta)\mathbf{Z}^\mathsf{T}\mathbf{X}\\ \mathbf{X}^\mathsf{T}\mathbf{Z}\Lambda(\theta) & \mathbf{X}^\mathsf{T}\mathbf{X}\end{bmatrix} \begin{bmatrix}\tilde{\mathbf{u}}(\theta)\\ \tilde{\beta}(\theta)\end{bmatrix} = \begin{bmatrix}\Lambda^\mathsf{T}(\theta)\mathbf{Z}^\mathsf{T}\mathbf{y}\\ \mathbf{X}^\mathsf{T}\mathbf{y}\end{bmatrix} \tag{4.41}$$

An effective way of determining the solution to a sparse, symmetric, positive definite system of equations such as (4.41) is the sparse Cholesky decomposition [Davis, 2006]. If $\mathbf{A}$ is a sparse, symmetric positive definite matrix then the sparse Cholesky factor with fill-reducing permutation $\mathbf{P}$ is the lower-triangular matrix $\mathbf{L}$ such that

$$\mathbf{L}\mathbf{L}^\mathsf{T} = \mathbf{P}\mathbf{A}\mathbf{P}^\mathsf{T}. \tag{4.42}$$

(Technically, the factor $\mathbf{L}$ is only determined up to changes in the sign of the diagonal elements. By convention we require the diagonal elements to be positive.)

The fill-reducing permutation represented by the permutation matrix $\mathbf{P}$, which is determined from the pattern of nonzeros in $\mathbf{A}$ but does not depend on particular values of those nonzeros, can have a profound impact on the number of nonzeros in $\mathbf{L}$ and hence on the speed with which $\mathbf{L}$ can be calculated from $\mathbf{A}$.

In most applications of linear mixed models the matrix $\mathbf{Z}\Lambda(\theta)$ is sparse while $\mathbf{X}$ is dense or close to it so the permutation matrix $\mathbf{P}$ can be restricted to the form

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_\mathbf{Z} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_\mathbf{X} \end{bmatrix} \tag{4.43}$$

without loss of efficiency. In fact, in most cases we can set $\mathbf{P}_\mathbf{X} = \mathbf{I}_p$ without loss of efficiency.

Let us assume that the permutation matrix is required to be of the form (4.43) so that we can write the Cholesky factorization for the positive definite system (4.41) as

$$\begin{bmatrix} \mathbf{L}_\mathbf{Z} & \mathbf{0} \\ \mathbf{L}_{\mathbf{XZ}} & \mathbf{L}_\mathbf{X} \end{bmatrix} \begin{bmatrix} \mathbf{L}_\mathbf{Z} & \mathbf{0} \\ \mathbf{L}_{\mathbf{XZ}} & \mathbf{L}_\mathbf{X} \end{bmatrix}^\mathsf{T} = \begin{bmatrix} \mathbf{P}_\mathbf{Z} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_\mathbf{X} \end{bmatrix} \begin{bmatrix} \Lambda^\mathsf{T}(\theta)\mathbf{Z}^\mathsf{T}\mathbf{Z}\Lambda(\theta) + \mathbf{I}_q & \Lambda^\mathsf{T}(\theta)\mathbf{Z}^\mathsf{T}\mathbf{X} \\ \mathbf{X}^\mathsf{T}\mathbf{Z}\Lambda(\theta) & \mathbf{X}^\mathsf{T}\mathbf{X} \end{bmatrix} \begin{bmatrix} \mathbf{P}_\mathbf{Z} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_\mathbf{X} \end{bmatrix}^\mathsf{T}. \tag{4.44}$$

The discrepancy can now be written in the canonical form

$$d(\mathbf{u}|\mathbf{y},\theta,\beta) = \tilde{d}(\mathbf{y},\theta) + \left\| \begin{bmatrix} \mathbf{L}_\mathbf{Z}^\mathsf{T} & \mathbf{L}_{\mathbf{XZ}}^\mathsf{T} \\ \mathbf{0} & \mathbf{L}_\mathbf{X}^\mathsf{T} \end{bmatrix} \begin{bmatrix} \mathbf{P}_\mathbf{Z}(\mathbf{u}-\tilde{\mathbf{u}}) \\ \mathbf{P}_\mathbf{X}(\beta-\tilde{\beta}) \end{bmatrix} \right\|^2 \tag{4.45}$$

where

$$\tilde{d}(\mathbf{y},\theta) = d(\tilde{\mathbf{u}}(\theta)|\mathbf{y},\theta,\tilde{\beta}(\theta)) \tag{4.46}$$

is the minimum discrepancy, given θ.

4.4.2 The profiled likelihood for linear mixed models

Substituting (4.45) into (4.38) provides the unnormalized conditional density $h(\mathbf{u}|\mathbf{y},\theta,\beta,\sigma)$ on the deviance scale as

$$-2\log(h(\mathbf{u}|\mathbf{y},\theta,\beta,\sigma))$$
$$= (n+q)\log(2\pi\sigma^2) + \frac{\tilde{d}(\mathbf{y},\theta) + \left\| \begin{bmatrix} \mathbf{L}_{\mathbf{Z}}^{\mathsf{T}} & \mathbf{L}_{\mathbf{XZ}}^{\mathsf{T}} \\ \mathbf{0} & \mathbf{L}_{\mathbf{X}}^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} \mathbf{P}_{\mathbf{Z}}(\mathbf{u}-\tilde{\mathbf{u}}) \\ \mathbf{P}_{\mathbf{X}}(\beta-\tilde{\beta}) \end{bmatrix} \right\|^2}{\sigma^2}. \quad (4.47)$$

As shown in Appendix **??**, the integral of a quadratic form on the deviance scale, such as (4.47), is easily evaluated, providing the log-likelihood, $\ell(\theta,\beta,\sigma|\mathbf{y})$, as

$$-2\ell(\theta,\beta,\sigma|\mathbf{y})$$
$$= -2\log\left(L(\theta,\beta,\sigma|\mathbf{y})\right)$$
$$= n\log(2\pi\sigma^2) + \log(|\mathbf{L}_{\mathbf{Z}}|^2) + \frac{\tilde{d}(\mathbf{y},\theta) + \left\|\mathbf{L}_{\mathbf{X}}^{\mathsf{T}}\mathbf{P}_{\mathbf{X}}(\beta-\tilde{\beta})\right\|^2}{\sigma^2}, \quad (4.48)$$

from which we can see that the conditional estimate of β, given θ, is $\tilde{\beta}(\theta)$ and the conditional estimate of σ, given θ, is

$$\tilde{\sigma^2}(\theta) = \frac{\tilde{d}(\theta|\mathbf{y})}{n}. \quad (4.49)$$

Substituting these conditional estimates into (4.48) produces the *profiled likelihood*, $\tilde{L}(\theta|\mathbf{y})$, as

$$-2\tilde{\ell}(\theta|\mathbf{y})) = \log(|\mathbf{L}_{\mathbf{Z}}(\theta)|^2) + n\left(1+\log\left(\frac{2\pi\tilde{d}(\mathbf{y},\theta)}{n}\right)\right). \quad (4.50)$$

The maximum likelihood estimate of θ can then be expressed as

$$\widehat{\theta}_L = \arg\min_{\theta}\left(-2\tilde{\ell}(\theta|\mathbf{y})\right). \quad (4.51)$$

from which the ML estimates of σ^2 and β are evaluated as

$$\widehat{\sigma_L^2} = \frac{\tilde{d}(\widehat{\theta}_L,\mathbf{y})}{n} \quad (4.52)$$
$$\widehat{\beta}_L = \tilde{\beta}(\widehat{\theta}_L). \quad (4.53)$$

The important thing to note about optimizing the profiled likelihood, (4.50), is that it is a m-dimensional optimization problem and typically m is very small.

4.4.3 The REML criterion

In practice the so-called REML estimates of variance components are often preferred to the maximum likelihood estimates. ("REML" can be considered to be an acronym for "restricted" or "residual" maximum likelihood, although neither term is completely accurate because these estimates do not maximize a likelihood.) We can motivate the use of the REML criterion by considering a linear regression model,

$$\mathscr{Y} \sim \mathscr{N}(\mathbf{X}\beta, \sigma^2\mathbf{I}_n), \tag{4.54}$$

in which we typically estimate σ^2 by

$$\widehat{\sigma^2_R} = \frac{\|\mathbf{y} - \mathbf{X}\widehat{\beta}\|^2}{n-p} \tag{4.55}$$

even though the maximum likelihood estimate of σ^2 is

$$\widehat{\sigma^2_L} = \frac{\|\mathbf{y} - \mathbf{X}\widehat{\beta}\|^2}{n}. \tag{4.56}$$

The argument for preferring $\widehat{\sigma^2_R}$ to $\widehat{\sigma^2_L}$ as an estimate of σ^2 is that the numerator in both estimates is the sum of squared residuals at $\widehat{\beta}$ and, although the residual vector $\mathbf{y} - \mathbf{X}\beta$ is an n-dimensional vector, the residual at $\widehat{\theta}$ satisfies p linearly independent constraints, $\mathbf{X}^\mathsf{T}(\mathbf{y} - \mathbf{X}\widehat{\beta}) = \mathbf{0}$. That is, the residual at $\widehat{\theta}$ is the projection of the observed response vector, $\mathbf{y}$, into an $(n-p)$-dimensional linear subspace of the n-dimensional response space. The estimate $\widehat{\sigma^2_R}$ takes into account the fact that σ^2 is estimated from residuals that have only $n-p$ *degrees of freedom*.

The REML criterion for determining parameter estimates $\widehat{\theta}_R$ and $\widehat{\sigma^2_R}$ in a linear mixed model has the property that these estimates would specialize to $\widehat{\sigma^2_R}$ from (4.55) for a linear regression model. Although not usually derived in this way, the REML criterion can be expressed as

$$c_R(\theta, \sigma|\mathbf{y}) = -2\log \int_{\mathbb{R}^p} L(\mathbf{u}|\mathbf{y}, \theta, \beta, \sigma)\, d\beta \tag{4.57}$$

on the deviance scale. The REML estimates $\widehat{\theta}_R$ and $\widehat{\sigma^2_R}$ minimize $c_R(\theta, \sigma|\mathbf{y})$.

The profiled REML criterion, a function of θ only, is

$$\tilde{c}_R(\theta|\mathbf{y}) = \log(|\mathbf{L}_\mathbf{Z}(\theta)|^2 |\mathbf{L}_\mathbf{X}(\theta)|^2) + (n-p)\left(1 + \log\left(\frac{2\pi\tilde{d}(\theta|\mathbf{y})}{n-p}\right)\right) \tag{4.58}$$

and the REML estimate of θ is

$$\widehat{\theta}_R = \arg\min_{\theta} \tilde{c}_R(\theta, \mathbf{y}). \tag{4.59}$$

The REML estimate of σ^2 is $\widehat{\sigma^2_R} = \tilde{d}(\widehat{\theta}_R|\mathbf{y})/(n-p)$.

It is not entirely clear how one would define a "REML estimate" of β because the REML criterion, $c_R(\theta, \sigma|\mathbf{y})$, defined in (4.57), does not depend on β. However, it is customary (and not unreasonable) to use $\widehat{\beta}_R = \tilde{\beta}(\widehat{\theta}_R)$ as the REML estimate of β.

Note that the profiled REML criterion can be evaluated from a sparse Cholesky decomposition like that in (4.44) but without the requirement that the permutation can be applied to the columns of $\mathbf{Z}\Lambda(\theta)$ separately from the columnns of $\mathbf{X}$. That is, we can use a general fill-reducing permutation rather than the specific form (4.43) with separate permutations represented by $\mathbf{P_Z}$ and $\mathbf{P_X}$. This can be useful in cases where both $\mathbf{Z}$ and $\mathbf{X}$ are large and sparse.

4.4.4 Summary for linear mixed models

A linear mixed model is characterized by the conditional distribution

$$(\mathscr{Y}|\mathscr{U} = \mathbf{u}) \sim \mathscr{N}(\gamma(\mathbf{u}, \theta, \beta), \sigma^2\mathbf{I}_n) \text{ where } \gamma(\mathbf{u}, \theta, \beta) = \mathbf{Z}\Lambda(\theta)\mathbf{u} + \mathbf{X}\beta \tag{4.60}$$

and the unconditional distribution $\mathscr{U} \sim \mathscr{N}(\mathbf{0}, \sigma^2\mathbf{I}_q)$. The discrepancy function,

$$d(\mathbf{u}|\mathbf{y}, \theta, \beta) = \|\mathbf{y} - \gamma(\mathbf{u}, \theta, \beta)\|^2 + \|\mathbf{u}\|^2,$$

is minimized at the conditional mode, $\tilde{\mathbf{u}}(\theta)$, and the conditional estimate, $\tilde{\beta}(\theta)$, which are the solutions to the sparse, positive-definite linear system

$$\begin{bmatrix} \Lambda^\mathsf{T}(\theta)\mathbf{Z}^\mathsf{T}\mathbf{Z}\Lambda(\theta) + \mathbf{I}_q & \Lambda^\mathsf{T}(\theta)\mathbf{Z}^\mathsf{T}\mathbf{X} \\ \mathbf{X}^\mathsf{T}\mathbf{Z}\Lambda(\theta) & \mathbf{X}^\mathsf{T}\mathbf{X} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{u}}(\theta) \\ \tilde{\beta}(\theta) \end{bmatrix} = \begin{bmatrix} \Lambda^\mathsf{T}(\theta)\mathbf{Z}^\mathsf{T}\mathbf{y} \\ \mathbf{X}^\mathsf{T}\mathbf{y} \end{bmatrix}.$$

In the process of solving this system we create the sparse left Cholesky factor, $L_\mathbf{Z}(\theta)$, which is a lower triangular sparse matrix satisfying

$$\mathbf{L_Z}(\theta)\mathbf{L_Z}(\theta)^\mathsf{T} = \mathbf{P_Z}\left(\Lambda^\mathsf{T}(\theta)\mathbf{Z}^\mathsf{T}\mathbf{Z}\Lambda(\theta) + \mathbf{I}_q\right)\mathbf{P}_\mathbf{Z}^\mathsf{T}$$

where $\mathbf{P_Z}$ is a permutation matrix representing a fill-reducing permutation formed from the pattern of nonzeros in $\mathbf{Z}\Lambda(\theta)$ for any θ not on the boundary of the parameter region. (The values of the nonzeros depend on θ but the pattern doesn't.)

The profiled log-likelihood, $\tilde{\ell}(\theta|\mathbf{y})$, is

$$-2\tilde{\ell}(\theta|\mathbf{y}) = \log(|\mathbf{L_Z}(\theta)|^2) + n\left(1 + \log\left(\frac{2\pi\tilde{d}(\mathbf{y}, \theta)}{n}\right)\right)$$

where $\tilde{d}(\mathbf{y},\theta) = d(\tilde{\mathbf{u}}(\theta)|\mathbf{y},\tilde{\beta}(\theta),\theta)$.

4.5 Generalizing the discrepancy function

Because one of the factors influencing the choice of implementation for linear mixed models is the extent to which the methods can also be applied to other mixed models, we describe several other classes of mixed models before discussing the implementation details for linear mixed models. At the core of our methods for determining the maximum likelihood estimates (MLEs) of the parameters in the mixed model are methods for minimizing the discrepancy function with respect to the coefficients $\mathbf{u}$ and β in the linear predictor $\gamma(\mathbf{u},\theta,\beta)$.

In this section we describe the general form of the discrepancy function that we will use and a penalized iteratively reweighted least squares (PIRLS) algorithm for determining the conditional modes $\tilde{\mathbf{u}}(\theta)$ and $\tilde{\beta}(\theta)$. We then describe several types of mixed models and the form of the discrepancy function for each.

4.5.1 A weighted residual sum of squares

As shown in §4.4.1, the discrepancy function for a linear mixed model has the form of a penalized residual sum of squares from a linear model (4.40). In this section we generalize that definition to

$$d(\mathbf{u}|\mathbf{y},\theta,\beta) = \left\|\mathbf{W}^{1/2}(\mu)\left[\mathbf{y}-\mu_{\mathscr{Y}|\mathscr{U}}(\mathbf{u},\theta,\beta)\right]\right\|^2 + \|\mathbf{0}-\mathbf{u}\|^2. \tag{4.61}$$

where $\mathbf{W}$ is an $n\times n$ diagonal matrix, called the *weights matrix*, with positive diagonal elements and $\mathbf{W}^{1/2}$ is the diagonal matrix with the square roots of the weights on the diagonal. The ith weight is inversely proportional to the conditional variances of $(\mathscr{Y}|\mathscr{U}=\mathbf{u})$ and may depend on the conditional mean, $\mu_{\mathscr{Y}|\mathscr{U}}$.

We allow the conditional mean to be a nonlinear function of the linear predictor, but with certain restrictions. We require that the mapping from $\mathbf{u}$ to $\mu_{\mathscr{Y}|\mathscr{U}=\mathbf{u}}$ be expressed as

$$\mathbf{u} \to \gamma \to \eta \to \mu \tag{4.62}$$

where $\gamma = \mathbf{Z}\Lambda(\theta)\mathbf{u}+\mathbf{X}\theta$ is an ns-dimensional vector ($s>0$) while η and μ are n-dimensional vectors.

The map $\eta\to\mu$ has the property that μ_i depends only on η_i, $i=1,\ldots,n$. The map $\gamma\to\eta$ has a similar property in that, if we write γ as an $n\times s$ matrix

■ such that

$$\gamma = \blacksquare \tag{4.63}$$

(i.e. concatenating the columns of ■ produces γ) then η_i depends only on the ith row of ■, $i = 1,\ldots,n$. Thus the Jacobian matrix $\frac{d\mu}{d\eta^\mathsf{T}}$ is an $n \times n$ diagonal matrix and the Jacobian matrix $\frac{d\eta}{d\gamma^\mathsf{T}}$ is the horizontal concatenation of s diagonal $n \times n$ matrices.

For historical reasons, the function that maps η_i to μ_i is called the *inverse link* function and is written $\mu = g^{-1}(\eta)$. The *link function*, naturally, is $\eta = g(\mu)$. When applied component-wise to vectors μ or η we write these as $\eta = \mathbf{g}(\mu)$ and $\mu = \mathbf{g}^{-1}(\eta)$.

Recall that the conditional distribution, $(\mathscr{Y}_i|\mathscr{U} = \mathbf{u})$, is required to be independent of $(\mathscr{Y}_j|\mathscr{U} = \mathbf{u})$ for $i, j = 1,\ldots,n, i \neq j$ and that all the component conditional distributions must be of the same form and differ only according to the value of the conditional mean.

Depending on the family of the conditional distributions, the allowable values of the μ_i may be in a restricted range. For example, if the conditional distributions are Bernoulli then $0 \leq \mu_i \leq 1, i = 1,\ldots,n$. If the conditional distributions are Poisson then $0 \leq \mu_i, i = 1,\ldots,n$. A characteristic of the link function, g, is that it must map the restricted range to an unrestricted range. That is, a link function for the Bernoulli distribution must map $[0,1]$ to $[-\infty,\infty]$ and must be invertible within the range.

The mapping from γ to η is defined by a function $m : \mathbb{R}^s \to \mathbb{R}$, called the *nonlinear model* function, such that $\eta_i = m(\gamma_i), i = 1,\ldots,n$ where γ_i is the ith row of ■. The vector-valued function is $\eta = \mathbf{m}(\gamma)$.

Determining the conditional modes, $\tilde{\mathbf{u}}(\mathbf{y}|\theta)$, and $\tilde{\beta}(\mathbf{y}|\theta)$, that jointly minimize the discrepancy,

$$\begin{bmatrix}\tilde{\mathbf{u}}(\mathbf{y}|\theta)\\ \tilde{\beta}(\mathbf{y}|\theta)\end{bmatrix} = \arg\min_{\mathbf{u},\beta}\left[(\mathbf{y}-\mu)^\mathsf{T}\mathbf{W}(\mathbf{y}-\mu) + \|\mathbf{u}\|^2\right] \tag{4.64}$$

becomes a weighted, nonlinear least squares problem except that the weights, $\mathbf{W}$, can depend on μ and, hence, on $\mathbf{u}$ and β.

In describing an algorithm for linear mixed models we called $\tilde{\beta}(\theta)$ the *conditional estimate*. That name reflects that fact that this is the maximum likelihood estimate of β for that particular value of θ. Once we have determined the MLE, $\widehat{(\theta)}_L$ of θ, we have a "plug-in" estimator, $\widehat{\beta}_L = \tilde{\beta}(\theta)$ for β.

This property does not carry over exactly to other forms of mixed models. The values $\tilde{\mathbf{u}}(\theta)$ and $\tilde{\beta}(\theta)$ are conditional modes in the sense that they are the coefficients in γ that jointly maximize the unscaled conditional density $h(\mathbf{u}|\mathbf{y},\theta,\beta,\sigma)$. Here we are using the adjective "conditional" more in the sense of conditioning on $\mathscr{Y} = \mathbf{y}$ than in the sense of conditioning on θ, although these values are determined for a fixed value of θ.

4.5.2 The PIRLS algorithm for $\tilde{\mathbf{u}}$ and $\tilde{\beta}$

The penalized, iteratively reweighted, least squares (PIRLS) algorithm to determine $\tilde{\mathbf{u}}(\theta)$ and $\tilde{\beta}(\theta)$ is a form of the Fisher scoring algorithm. We fix the weights matrix, $\mathbf{W}$, and use penalized, weighted, nonlinear least squares to minimize the penalized, weighted residual sum of squares conditional on these weights. Then we update the weights to those determined by the current value of μ and iterate.

To describe this algorithm in more detail we will use parenthesized superscripts to denote the iteration number. Thus $\mathbf{u}^{(0)}$ and $\beta^{(0)}$ are the initial values of these parameters, while $\mathbf{u}^{(i)}$ and $\beta^{(i)}$ are the values at the ith iteration. Similarly $\gamma^{(i)} = \mathbf{Z}\Lambda(\theta)\mathbf{u}^{(i)} + \mathbf{X}\beta^{(i)}$, $\eta^{(i)} = \mathbf{m}(\gamma^{(i)})$ and $\mu^{(i)} = \mathbf{g}^{-1}(\eta^{(i)})$.

We use a penalized version of the Gauss-Newton algorithm [Bates and Watts, 1988, ch. 2] for which we define the weighted Jacobian matrices

$$\mathbf{U}^{(i)} = \mathbf{W}^{1/2}\left.\frac{d\mu}{d\mathbf{u}^\mathsf{T}}\right|_{\mathbf{u}=\mathbf{u}^{(i)},\beta=\beta^{(i)}} = \mathbf{W}^{1/2}\left.\frac{d\mu}{d\eta^\mathsf{T}}\right|_{\eta^{(i)}}\left.\frac{d\eta}{d\gamma^\mathsf{T}}\right|_{\gamma^{(i)}}\mathbf{Z}\Lambda(\theta) \tag{4.65}$$

$$\mathbf{V}^{(i)} = \mathbf{W}^{1/2}\left.\frac{d\mu}{d\beta^\mathsf{T}}\right|_{\mathbf{u}=\mathbf{u}^{(i)},\beta=\beta^{(i)}} = \mathbf{W}^{1/2}\left.\frac{d\mu}{d\eta^\mathsf{T}}\right|_{\eta^{(i)}}\left.\frac{d\eta}{d\gamma^\mathsf{T}}\right|_{\gamma^{(i)}}\mathbf{X} \tag{4.66}$$

of dimension $n\times q$ and $n\times p$, respectively. The increments at the ith iteration, $\delta_{\mathbf{u}}^{(i)}$ and $\delta_{\beta}^{(i)}$, are the solutions to

$$\begin{bmatrix}\mathbf{U}^{(i)\mathsf{T}}\mathbf{U}^{(i)}+\mathbf{I}_q & \mathbf{U}^{(i)\mathsf{T}}\mathbf{V}^{(i)}\\ \mathbf{V}^{(i)\mathsf{T}}\mathbf{U}^{(i)} & \mathbf{V}^{(i)\mathsf{T}}\mathbf{V}^{(i)}\end{bmatrix}\begin{bmatrix}\delta_{\mathbf{u}}^{(i)}\\ \delta_{\beta}^{(i)}\end{bmatrix} = \begin{bmatrix}\mathbf{U}^{(i)\mathsf{T}}\mathbf{W}^{1/2}(\mathbf{y}-\mu^{(i)})-\mathbf{u}^{(i)}\\ \mathbf{U}^{(i)\mathsf{T}}\mathbf{W}^{1/2}(\mathbf{y}-\mu^{(i)})\end{bmatrix} \tag{4.67}$$

providing the updated parameter values

$$\begin{bmatrix}\mathbf{u}^{(i+1)}\\ \beta^{(i+1)}\end{bmatrix} = \begin{bmatrix}\mathbf{u}^{(i)}\\ \beta^{(i)}\end{bmatrix} + \lambda\begin{bmatrix}\delta_{\mathbf{u}}^{(i)}\\ \delta_{\beta}^{(i)}\end{bmatrix} \tag{4.68}$$

where $\lambda > 0$ is a step factor chosen to ensure that

$$(\mathbf{y}-\mu^{(i+1)})^\mathsf{T}\mathbf{W}(\mathbf{y}-\mu^{(i+1)}) + \|\mathbf{u}^{(i+1)}\|^2 < (\mathbf{y}-\mu^{(i)})^\mathsf{T}\mathbf{W}(\mathbf{y}-\mu^{(i)}) + \|\mathbf{u}^{(i)}\|^2. \tag{4.69}$$

In the process of solving for the increments we form the sparse, lower triangular, Cholesky factor, $\mathbf{L}^{(i)}$, satisfying

$$\mathbf{L}^{(i)}\mathbf{L}^{(i)\mathsf{T}} = \mathbf{P}_\mathbf{Z}\left(\mathbf{U}^{(i)\mathsf{T}}\mathbf{U}^{(i)}+\mathbf{I}_n\right)\mathbf{P}_\mathbf{Z}^\mathsf{T}. \tag{4.70}$$

After each successful iteration, determining new values of the coefficients, $\mathbf{u}^{(i+1)}$ and $\beta^{(i+1)}$, that reduce the penalized, weighted residual sum of squqres, we update the weights matrix to $\mathbf{W}(\mu^{(i+1)})$ and the weighted Jacobians, $\mathbf{U}^{(i+1)}$ and $\mathbf{V}^{(i+1)}$, then iterate. Convergence is determined according to the orthogo-

nality convergence criterion [Bates and Watts, 1988, ch. 2], suitably adjusted for the weights matrix and the penalty.

4.5.3 Weighted linear mixed models

One of the simplest generalizations of linear mixed models is a weighted linear mixed model where $s = 1$, the link function, g, and the nonlinear model function, m, are both the identity, the weights matrix, $\mathbf{W}$, is constant and the conditional distribution family is Gaussian. That is, the conditional distribution can be written

$$(\mathscr{Y}|\mathscr{U} = \mathbf{u}) \sim \mathscr{N}(\gamma(\mathbf{u}, \theta, \beta), \sigma^2\mathbf{W}^{-1}) \tag{4.71}$$

with discrepancy function

$$d(\mathbf{u}|\mathbf{y}, \theta, \beta) = \left\|\mathbf{W}^{1/2}(\mathbf{y} - \mathbf{Z}\Lambda(\theta)\mathbf{u} - \mathbf{X}\theta)\right\|^2 + \|\mathbf{u}\|^2. \tag{4.72}$$

The conditional mode, $\tilde{\mathbf{u}}(\theta)$, and the conditional estimate, $\tilde{\beta}(\theta)$, are the solutions to

$$\begin{bmatrix} \Lambda^\mathsf{T}(\theta)\mathbf{Z}^\mathsf{T}\mathbf{W}\mathbf{Z}\Lambda(\theta) + \mathbf{I}_q & \Lambda^\mathsf{T}(\theta)\mathbf{Z}^\mathsf{T}\mathbf{W}\mathbf{X} \\ \mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{Z}\Lambda(\theta) & \mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{u}}(\theta) \\ \tilde{\beta}(\theta) \end{bmatrix} = \begin{bmatrix} \Lambda^\mathsf{T}(\theta)\mathbf{Z}^\mathsf{T}\mathbf{W}\mathbf{y} \\ \mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{y} \end{bmatrix}, \tag{4.73}$$

which can be solved directly, and the Cholesky factor, $\mathbf{L}_\mathbf{Z}(\theta)$, satisfies

$$\mathbf{L}_\mathbf{Z}(\theta)\mathbf{L}_\mathbf{Z}(\theta)^\mathsf{T} = \mathbf{P}_\mathbf{Z}\left(\Lambda^\mathsf{T}(\theta)\mathbf{Z}^\mathsf{T}\mathbf{W}\mathbf{Z}\Lambda(\theta) + \mathbf{I}_q\right)\mathbf{P}_\mathbf{Z}^\mathsf{T}. \tag{4.74}$$

The profiled log-likelihood, $\tilde{\ell}(\theta|\mathbf{y})$, is

$$-2\tilde{\ell}(\theta|\mathbf{y}) = \log\left(\frac{|\mathbf{L}_\mathbf{Z}(\theta)|^2}{|\mathbf{W}|}\right) + n\left(1 + \log\left(\frac{2\pi\tilde{d}(\mathbf{y}, \theta)}{n}\right)\right). \tag{4.75}$$

If the matrix $\mathbf{W}$ is fixed then we can ignore the term $|\mathbf{W}|$ in (4.75) when determining the MLE, $\widehat{\theta}_L$. However, in some models, we use a parameterized weight matrix, $\mathbf{W}(\phi)$, and wish to determine the MLEs, $\widehat{\phi}_L$ and $\widehat{\theta}_L$ simultaneously. In these cases we must include the term involving $|\mathbf{W}(\phi)|$ when evaluating the profiled log-likelihood.

Note that we must define the parameterization of $\mathbf{W}(\phi)$ such that σ^2 and ϕ are not a redundant parameterization of $\sigma^2\mathbf{W}(\phi)$. For example, we could require that the first diagonal element of $\mathbf{W}$ be unity.

4.5.4 Nonlinear mixed models

In an unweighted, nonlinear mixed model the conditional distribution is Gaussian, the link, g, is the identity and the weights matrix, $\mathbf{W} = \mathbf{I}_n$. That is,

$$(\mathscr{Y}|\mathscr{U} = \mathbf{u}) \sim \mathscr{N}(\mathbf{m}(\gamma), \sigma^2\mathbf{I}_n) \tag{4.76}$$

with discrepancy function

$$d(\mathbf{u}|\mathbf{y}, \theta, \beta) = \|\mathbf{y} - \mu\|^2 + \|\mathbf{u}\|^2. \tag{4.77}$$

For a given value of θ we determine the conditional modes, $\tilde{\mathbf{u}}(\theta)$ and $\tilde{\beta}(\theta)$, as the solution to the penalized nonlinear least squares problem

$$\begin{bmatrix}\tilde{\mathbf{u}}(\theta)\\ \tilde{\beta}(\theta)\end{bmatrix} = \arg\min_{\mathbf{u},\theta} d(\mathbf{u}|\mathbf{y}, \theta, \beta) \tag{4.78}$$

and we write the minimum discrepancy, given $\mathbf{y}$ and θ, as

$$\tilde{d}(\mathbf{y}, \theta) = d(\tilde{\mathbf{u}}(\theta)|\mathbf{y}, \theta, \tilde{\beta}(\theta)). \tag{4.79}$$

Let $\tilde{\mathbf{L}}_Z(\theta)$ and $\tilde{\mathbf{L}}_X(\theta)$ be the Cholesky factors at θ, $\tilde{\beta}(\theta)$ and $\tilde{\mathbf{u}}(\theta)$. Then the *Laplace approximation* to the log-likelihood is

$$-2\ell_P(\theta, \beta, \sigma|\mathbf{y}) \approx n\log(2\pi\sigma^2) + \log(|\tilde{\mathbf{L}}_\mathbf{Z}|^2) + \frac{\tilde{d}(\mathbf{y}, \theta) + \left\|\tilde{\mathbf{L}}_\mathbf{X}^\mathsf{T}(\beta - \tilde{\beta})\right\|^2}{\sigma^2}, \tag{4.80}$$

producing the approximate profiled log-likelihood, $\tilde{\ell}_P(\theta|\mathbf{y})$,

$$-2\tilde{\ell}_P(\theta|\mathbf{y}) \approx \log(|\tilde{\mathbf{L}}_\mathbf{Z}|^2) + n\left(1 + \log(2\pi\tilde{d}(\mathbf{y}, \theta)/n)\right). \tag{4.81}$$

4.5.4.1 Nonlinear mixed model summary

In a nonlinear mixed model we determine the parameter estimate, $\widehat{\theta}_P$, from the Laplace approximation to the log-likelihood as

$$\widehat{\theta}_P = \arg\max_\theta \tilde{\ell}_P(\theta|\mathbf{y}) = \arg\min_\theta \log(|\tilde{\mathbf{L}}_\mathbf{Z}|^2) + n\left(1 + \log(2\pi\tilde{d}(\mathbf{y}, \theta)/n)\right). \tag{4.82}$$

Each evaluation of $\tilde{\ell}_P(\theta|\mathbf{y})$ requires a solving the penalized nonlinear least squares problem (4.78) simultaneously with respect to both sets of coefficients, $\mathbf{u}$ and β, in the linear predictor, γ.

For a weighted nonlinear mixed model with fixed weights, $\mathbf{W}$, we replace the unweighted discrepancy function $d(\mathbf{u}|\mathbf{y}, \theta, \beta)$ with the weighted discrepancy function,

4.6 Details of the implementation

4.6.1 Implementation details for linear mixed models

The crucial step in implementing algorithms for determining ML or REML estimates of the parameters in a linear mixed model is evaluating the factorization (4.44) for any θ satisfying $\theta_L \le \theta \le \theta_U$. We will assume that $\mathbf{Z}$ is sparse as is $\mathbf{Z}\Lambda(\theta)$.

When $\mathbf{X}$ is not sparse we will use the factorization (4.44) setting $\mathbf{P_X} = \mathbf{I}_p$ and storing $\mathbf{L_{XZ}}$ and $\mathbf{L_X}$ as dense matrices. The permutation matrix $\mathbf{P_Z}$ is determined from the pattern of non-zeros in $\mathbf{Z}\Lambda(\theta)$ which is does not depend on θ, as long as θ is not on the boundary. In fact, in most cases the pattern of non-zeros in $\mathbf{Z}\Lambda(\theta)$ is the same as the pattern of non-zeros in $\mathbf{Z}$. For many models, in particular models with scalar random effects (described later), the matrix $\Lambda(\theta)$ is diagonal.

Given a value of θ we determine the Cholesky factor $\mathbf{L_Z}$ satisfying

$$\mathbf{L_Z}\mathbf{L_Z}^\mathsf{T} = \mathbf{P_Z}(\Lambda^\mathsf{T}(\theta)\mathbf{Z}^\mathsf{T}\mathbf{Z}\Lambda(\theta) + \mathbf{I}_q)\mathbf{P_Z}^\mathsf{T}. \tag{4.83}$$

The CHOLMOD package allows for $\mathbf{L_Z}$ to be calculated directly from $\Lambda^\mathsf{T}(\theta)\mathbf{Z}^\mathsf{T}$ or from $\Lambda^\mathsf{T}(\theta)\mathbf{Z}^\mathsf{T}\mathbf{Z}\Lambda(\theta)$. The choice in implementation is whether to store $\mathbf{Z}^\mathsf{T}$ and update it to $\Lambda^\mathsf{T}(\theta)\mathbf{Z}$ or to store $\mathbf{Z}^\mathsf{T}\mathbf{Z}$ and use it to form $\Lambda^\mathsf{T}(\theta)\mathbf{Z}^\mathsf{T}\mathbf{Z}\Lambda(\theta)$ at each evaluation.

In the `lme4` package we store $\mathbf{Z}^\mathsf{T}$ and use it to form $\Lambda^\mathsf{T}(\theta)\mathbf{Z}^\mathsf{T}$ from which $\mathbf{L_Z}$ is evaluated. There are two reasons for this choice. First, the calculations for the more general forms of mixed models cannot be reduced to calculations involving $\mathbf{Z}^\mathsf{T}\mathbf{Z}$ and by expressing these calculations in terms of $\Lambda(\theta)\mathbf{Z}^\mathsf{T}$ for linear mixed models we can reuse the code for the more general models. Second, the calculation of $\Lambda(\theta)^\mathsf{T}\left(\mathbf{Z}^\mathsf{T}\mathbf{Z}\right)\Lambda(\theta)$ from $\mathbf{Z}^\mathsf{T}\mathbf{Z}$ is complicated compared to the calculation of $\Lambda(\theta)^\mathsf{T}\mathbf{Z}^\mathsf{T}$ from $\mathbf{Z}^\mathsf{T}$.

This choice is disadvantageous when $n \gg q$ because $\mathbf{Z}^\mathsf{T}$ is much larger than $\mathbf{Z}^\mathsf{T}\mathbf{Z}$, even when they are stored as sparse matrices. Evaluation of $\mathbf{L_Z}$ directly from $\mathbf{Z}^\mathsf{T}$ requires more storage and more calculation that evaluating $\mathbf{L_Z}$ from $\mathbf{Z}^\mathsf{T}\mathbf{Z}$.

Next we evaluate $\mathbf{L_{XZ}}^\mathsf{T}$ as the solution to

$$\mathbf{L_Z}\mathbf{L_{XZ}}^\mathsf{T} = \mathbf{P_Z}\Lambda^\mathsf{T}(\theta)\mathbf{Z}^\mathsf{T}\mathbf{X}. \tag{4.84}$$

Again we have the choice of calculating and storing $\mathbf{Z}^\mathsf{T}\mathbf{X}$ or storing $\mathbf{X}$ and using it to reevaluate $\mathbf{Z}^\mathsf{T}\mathbf{X}$. In the `lme4` package we store $\mathbf{X}$, because the calculations for the more general models cannot be expressed in terms of $\mathbf{Z}^\mathsf{T}\mathbf{X}$.

Finally $\mathbf{L_X}$ is evaluated as the (dense) solution to

$$\mathbf{L_X}\mathbf{L_X}^\mathsf{T} = \mathbf{X}^\mathsf{T}\mathbf{X} - \mathbf{L_{XZ}}\mathbf{L_{XZ}}. \tag{4.85}$$

from which $\tilde{\boldsymbol{\beta}}$ can be determined as the solution to dense system

$$\mathbf{L}_{\mathbf{X}}\mathbf{L}_{\mathbf{X}}\tilde{\boldsymbol{\beta}} = \mathbf{X}^{\mathsf{T}}\mathbf{y} \tag{4.86}$$

and $\tilde{\mathbf{u}}$ as the solution to the sparse system

$$\mathbf{L}_{\mathbf{Z}}\mathbf{L}_{\mathbf{Z}}\tilde{u} = \Lambda^{\mathsf{T}}\mathbf{Z}^{\mathsf{T}}\mathbf{y} \tag{4.87}$$

For many models, in particular models with scalar random effects, which are described later, the matrix $\Lambda(\theta)$ is diagonal. For such a model, if both $\mathbf{Z}$ and $\mathbf{X}$ are sparse and we plan to use the REML criterion then we create and store

$$\mathbf{A} = \begin{bmatrix} \mathbf{Z}^{\mathsf{T}}\mathbf{Z} & \mathbf{Z}^{\mathsf{T}}\mathbf{X} \\ \mathbf{X}^{\mathsf{T}}\mathbf{Z} & \mathbf{X}^{\mathsf{T}}\mathbf{X} \end{bmatrix} \quad \text{and} \quad \mathbf{c} = \begin{bmatrix} \mathbf{Z}^{\mathsf{T}}\mathbf{y} \\ \mathbf{X}^{\mathsf{T}}\mathbf{y} \end{bmatrix} \tag{4.88}$$

and determine a fill-reducing permutation, $\mathbf{P}$, for $\mathbf{A}$. Given a value of θ we create the factorization

$$\mathbf{L}(\theta)\mathbf{L}(\theta)^{\mathsf{T}} = \mathbf{P}\left(\begin{bmatrix} \Lambda(\theta) & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_p \end{bmatrix}\mathbf{A}\begin{bmatrix} \Lambda(\theta) & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_p \end{bmatrix} + \begin{bmatrix} \mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\right)\mathbf{P}^{\mathsf{T}} \tag{4.89}$$

solve for $\tilde{\mathbf{u}}(\theta)$ and $\tilde{\boldsymbol{\beta}}(\theta)$ in

$$\mathbf{L}\mathbf{L}^{\mathsf{T}}\mathbf{P}\begin{bmatrix} \tilde{\mathbf{u}}(\theta) \\ \tilde{\boldsymbol{\beta}}(\theta) \end{bmatrix} = \mathbf{P}\begin{bmatrix} \Lambda(\theta) & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_p \end{bmatrix}\mathbf{c} \tag{4.90}$$

then evaluate $\tilde{d}(\mathbf{y}|\theta)$ and the profiled REML criterion as

$$\tilde{d}_R(\theta|\mathbf{y}) = \log(|\mathbf{L}(\theta)|^2) + (n-p)\left(1 + \log\left(\frac{2\pi\tilde{d}(\mathbf{y}|\theta)}{n-p}\right)\right). \tag{4.91}$$

References

Douglas M. Bates and Donald G. Watts. *Nonlinear Regression Analysis and Its Applications*. Wiley, Hoboken, NJ, 1988. ISBN 0-471-81643-4.

Gregory Belenky, Nancy J. Wesensten, David R. Thorne, Maria L. Thomas, Helen C. Sing, Daniel P. Redmond, Michael B. Russo, and Thomas J. Balkin. Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: a sleep dose-response study. *Journal of Sleep Research*, 12:1–12, 2003.

G.E.P. Box and G.C. Tiao. *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, MA, 1973.

Bill Cleveland. *Visualizing Data*. Hobart Press, Summit, NJ, 1993.

Owen L. Davies and Peter L. Goldsmith, editors. *Statistical Methods in Research and Production*. Hafner, 4th edition, 1972.

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

Tim Davis. *Direct Methods for Sparse Linear Systems*. SIAM, Philadelphia, PA, 2006.

Friedrich Leisch. Sweave: Dynamic generation of statistical reports using literate data analysis. In Wolfgang Härdle and Bernd Rönz, editors, *Compstat 2002 — Proceedings in Computational Statistics*, pages 575–580. Physica Verlag, Heidelberg, 2002. URL `http://www.stat.uni-muenchen.de/~leisch/Sweave`. ISBN 3-7908-1517-9.

José C. Pinheiro and Douglas M. Bates. *Mixed-effects Models in S and S-PLUS*. Springer, 2000.

J. Rasbash, W. Browne, H. Goldstein, M. Yang, and I. Plewis. *A User's Guide to MLwiN*. Multilevel Models Project, Institute of Education, University of London, London, 2000.

Stephen W. Raudenbush and Anthony S. Bryk. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage, 2nd edition, 2002. ISBN 0-7619-1904-X.

Y. Sakamoto, M. Ishiguro, and G. Kitagawa. *Akaike Information Criterion Statistics*. Reidel, Dordrecht, Holland, 1986.

Deepayan Sarkar. *Lattice: Multivariate Data Visualization with R*. Springer, 2008.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

Index