Chapter 5
Computational Methods for Mixed Models

In this chapter we describe some of the details of the computational methods for fitting linear mixed 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, $\gamma = \mathbf{X}\beta + \mathbf{Z}\mathbf{b} = \mathbf{Z}\Lambda_\theta\mathbf{u} + \mathbf{X}\beta$, 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.

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.

5.1 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, $\mathscr{Y}$. Equation (1.1) defines the unconditional distribution of $\mathscr{B}$ and the conditional distribution of $\mathscr{Y}$, given $\mathscr{B} = \mathbf{b}$, as multivariate Gaussian distributions of the form

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

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

$$\mathbf{b}^\mathsf{T}\Sigma_\theta\mathbf{b} \geq 0, \quad \forall \mathbf{b} \neq \mathbf{0}. \tag{5.1}$$

(The symbol $\forall$ denotes "for all".) The fact that Σ_θ is positive semidefinite does not guarantee that Σ_θ^{-1} exists. We would need a stronger property, $\mathbf{b}^\mathsf{T}\Sigma_\theta\mathbf{b} > 0, \forall \mathbf{b} \neq \mathbf{0}$, 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. And it can do so. 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.

As defined in (1.2) a relative covariance factor, Λ_θ, is any matrix that satisfies

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

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

$$\mathscr{B} = \Lambda_\theta\mathscr{U}. \tag{5.2}$$

Although it may seem more intuitive to write $\mathscr{U}$ as a linear transformation of $\mathscr{B}$, we cannot do that when Λ_θ is singular, which is why (5.2) is in the form shown.

We can easily verify that (5.2) 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,

$$\mathrm{E}[\mathscr{B}] = \Lambda_\theta\mathrm{E}[\mathscr{U}] = \Lambda_\theta\mathbf{0} = \mathbf{0} \tag{5.3}$$

and

$$\begin{aligned}\mathrm{Var}(\mathcal{B}) &= \mathrm{E}\left[(\mathcal{B}-\mathrm{E}[\mathcal{B}])(\mathcal{B}-\mathrm{E}[\mathcal{B}])^\mathsf{T}\right] = \mathrm{E}\left[\mathcal{B}\mathcal{B}^\mathsf{T}\right] \\ &= \mathrm{E}\left[\Lambda_\theta \mathcal{U}\mathcal{U}^\mathsf{T}\Lambda_\theta^\mathsf{T}\right] = \Lambda_\theta\,\mathrm{E}[\mathcal{U}\mathcal{U}^\mathsf{T}]\Lambda_\theta^\mathsf{T} = \Lambda_\theta\,\mathrm{Var}(\mathcal{U})\Lambda_\theta^\mathsf{T} \\ &= \Lambda_\theta\,\sigma^2\mathbf{I}_q\Lambda_\theta^\mathsf{T} = \sigma^2\Lambda_\theta\Lambda_\theta^\mathsf{T} = \Sigma_\theta\end{aligned} \tag{5.4}$$

and have the desired form.

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

$$\begin{aligned}(\mathcal{Y}|\mathcal{U}=\mathbf{u}) &\sim \mathcal{N}(\mathbf{Z}\Lambda_\theta\mathbf{u}+\mathbf{X}\beta,\sigma^2\mathbf{I}_n) \\ \mathcal{U} &\sim \mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_q).\end{aligned} \tag{5.5}$$

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

$$\gamma = \mathbf{Z}\Lambda_\theta\mathbf{u}+\mathbf{X}\beta \tag{5.6}$$

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

$$\mu = \mathrm{E}[\mathcal{Y}|\mathcal{U}=\mathbf{u}]. \tag{5.7}$$

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 matrix $\mathbf{Z}$ must be $n\times q$ and $\mathbf{X}$ must be $n\times p$.

5.2 The Conditional Distribution $(\mathcal{U}|\mathcal{Y}=\mathbf{y})$

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

The likelihood of the parameters, θ, β, and σ, given the observed data, $\mathbf{y}_{\mathrm{obs}}$, is the probability density of $\mathcal{Y}$, evaluated at $\mathbf{y}_{\mathrm{obs}}$. Although the numerical values of the probability density and the likelihood are identical, the interpretations of these functions are different. In the density we consider the parameters to be fixed and the value of $\mathbf{y}$ as varying. In the likelihood we consider $\mathbf{y}$ to be fixed at $\mathbf{y}_{\mathrm{obs}}$ and the parameters, θ, β and σ, as varying.

The natural approach for evaluating the likelihood is to determine the marginal distribution of $\mathcal{Y}$, which in this case amounts to determining the marginal density of $\mathcal{Y}$, and evaluate that density at $\mathbf{y}_{\mathrm{obs}}$. 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

$$f_{\mathscr{U}|\mathscr{Y}}(\mathbf{u}|\mathbf{y}_{\text{obs}}) = \frac{h(\mathbf{u})}{\int_{\mathbb{R}^q} h(\mathbf{u})\,d\mathbf{u}}. \tag{5.8}$$

In some theoretical developments the normalizing constant, which is the integral in the denominator of an expression like (5.8), is not of interest. Here it is of interest because the normalizing constant is exactly the likelihood that we wish to evaluate,

$$L(\theta,\beta,\sigma|\mathbf{y}_{\text{obs}}) = \int_{\mathbb{R}^q} h(\mathbf{u})\,d\mathbf{u}. \tag{5.9}$$

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.

5.3 Integrating $h(\mathbf{u})$ in the Linear Mixed Model

The integral defining the likelihood in (5.9) 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,

$$\tilde{\mathbf{u}} = \arg\max_{\mathbf{u}} f_{\mathscr{U}|\mathscr{Y}}(\mathbf{u}|\mathbf{y}_{\text{obs}}) = \arg\max_{\mathbf{u}} h(\mathbf{u}) = \arg\max_{\mathbf{u}} f_{\mathscr{Y}|\mathscr{U}}(\mathbf{y}_{\text{obs}}|\mathbf{u})\, f_{\mathscr{U}}(\mathbf{u}) \quad (5.10)$$

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

In general, the *mode* of a continuous distribution is the value of the random variable that maximizes the density. The value $\tilde{\mathbf{u}}$ is called the conditional mode of $\mathbf{u}$, given $\mathscr{Y} = \mathbf{y}_{\text{obs}}$, because $\tilde{\mathbf{u}}$ maximizes the conditional density of $\mathscr{U}$ given $\mathscr{Y} = \mathbf{y}_{\text{obs}}$. The location of the maximum can be determined by maximizing the unnormalized conditional density because $h(\mathbf{u})$ is just a constant multiple of $f_{\mathscr{U}|\mathscr{Y}}(\mathbf{u}|\mathbf{y}_{\text{obs}})$. The last part of (5.10) is simply a re-expression of $h(\mathbf{u})$ as the product of $f_{\mathscr{Y}|\mathscr{U}}(\mathbf{y}_{\text{obs}}|\mathbf{u})$ and $f_{\mathscr{U}}(\mathbf{u})$. For a linear mixed model these densities are

$$f_{\mathscr{Y}|\mathscr{U}}(\mathbf{y}|\mathbf{u}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\|\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\Lambda_\theta\mathbf{u}\|^2}{2\sigma^2}\right) \quad (5.11)$$

$$f_{\mathscr{U}}(\mathbf{u}) = \frac{1}{(2\pi\sigma^2)^{q/2}} \exp\left(-\frac{\|\mathbf{u}\|^2}{2\sigma^2}\right) \quad (5.12)$$

with product

$$h(\mathbf{u}) = \frac{1}{(2\pi\sigma^2)^{(n+q)/2}} \exp\left(-\frac{\|\mathbf{y}_{\text{obs}} - \mathbf{X}\beta - \mathbf{Z}\Lambda_\theta\mathbf{u}\|^2 + \|\mathbf{u}\|^2}{2\sigma^2}\right). \quad (5.13)$$

On the deviance scale we have

$$-2\log\left(h(\mathbf{u})\right) = (n+q)\log(2\pi\sigma^2) + \frac{\|\mathbf{y}_{\text{obs}} - \mathbf{X}\beta - \mathbf{Z}\Lambda_\theta\mathbf{u}\|^2 + \|\mathbf{u}\|^2}{\sigma^2}. \quad (5.14)$$

Because (5.14) describes the negative log density, $\tilde{\mathbf{u}}$ will be the value of $\mathbf{u}$ that minimizes the expression on the right of (5.14).

The only part of the right hand side of (5.14) that depends on $\mathbf{u}$ is the numerator of the second term. Thus

$$\tilde{\mathbf{u}} = \arg\min_{\mathbf{u}} \|\mathbf{y}_{\text{obs}} - \mathbf{X}\beta - \mathbf{Z}\Lambda_\theta\mathbf{u}\|^2 + \|\mathbf{u}\|^2. \quad (5.15)$$

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, $\|\mathbf{y}_{\text{obs}} - \mathbf{X}\beta - \mathbf{Z}\Lambda_\theta\mathbf{u}\|^2$, 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 (i.e. producing a small residual sum of squares) against simplicity of the model (small $\|\mathbf{u}\|$). We refer to this type of criterion as

a smoothing objective, in the sense that it seeks to smooth out the fitted response by reducing model complexity while still retaining reasonable fidelity to the observed data.

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

$$r^2_{\theta,\beta} = \min_{\mathbf{u}} \|\mathbf{y}_{\text{obs}} - \mathbf{X}\beta - \mathbf{Z}\Lambda_\theta \mathbf{u}\|^2 + \|\mathbf{u}\|^2. \tag{5.16}$$

Notice that β only enters the right hand side of (5.16) through the linear predictor expression. 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

$$r^2_{\theta} = \min_{\mathbf{u},\beta} \|\mathbf{y}_{\text{obs}} - \mathbf{X}\beta - \mathbf{Z}\Lambda_\theta \mathbf{u}\|^2 + \|\mathbf{u}\|^2. \tag{5.17}$$

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

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

One way of expressing a penalized least squares problem like (5.16) is by incorporating the penalty as "pseudo-data" in an ordinary least squares problem. We extend the "response vector", which is $\mathbf{y}_{\text{obs}} - \mathbf{X}\beta$ 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

$$\tilde{\mathbf{u}} = \arg\min_{\mathbf{u}} \left\| \begin{bmatrix} \mathbf{y}_{\text{obs}} - \mathbf{X}\beta \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{Z}\Lambda_\theta \\ \mathbf{I}_q \end{bmatrix} \mathbf{u} \right\|^2 \tag{5.18}$$

with a solution that satisfies

$$\left(\Lambda_\theta^\top \mathbf{Z}^\top \mathbf{Z}\Lambda_\theta + \mathbf{I}_q\right) \tilde{\mathbf{u}} = \Lambda_\theta^\top \mathbf{Z}^\top (\mathbf{y}_{\text{obs}} - \mathbf{X}\beta) \tag{5.19}$$

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

$$\mathbf{L}_\theta \mathbf{L}_\theta^\top = \Lambda_\theta^\top \mathbf{Z}^\top \mathbf{Z}\Lambda_\theta + \mathbf{I}_q. \tag{5.20}$$

The actual evaluation of sparse Cholesky factor, $\mathbf{L}_\theta$, often incorporates a *fill-reducing permutation*, which we describe next.

5.4.1 The Fill-reducing Permutation, P

In earlier chapters we have seen that 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, $\mathbf{Z}\Lambda_\theta$, 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 apply the permutation without ever constructing $\mathbf{P}$. That is, the matrix $\mathbf{P}$ is only a notational convenience only.

The matrix $\mathbf{P}$ consists of permuted columns of the identity matrix, $\mathbf{I}_q$, and it is easy to establish that the inverse permutation corresponds to multiplication by $\mathbf{P}^\mathsf{T}$. Because multiplication by $\mathbf{P}$ or by $\mathbf{P}^\mathsf{T}$ simply re-orders the components of a vector, the length of the vector is unchanged. Thus,

$$\|\mathbf{P}\mathbf{u}\|^2 = \|\mathbf{u}\|^2 = \|\mathbf{P}^\mathsf{T}\mathbf{u}\|^2 \tag{5.21}$$

and we can express the penalty in (5.17) in any of these three forms. The properties of $\mathbf{P}$ that it preserves lengths of vectors and that its transpose is its inverse are summarized by stating that $\mathbf{P}$ is an *orthogonal matrix*.

The permutation represented by $\mathbf{P}$ is determined from the structure of $\Lambda_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{Z}\Lambda_\theta + \mathbf{I}_q$ for some initial value of θ. The particular value of θ does not affect the result because the permutation depends only the positions of the non-zeros, not the numerical values at these positions.

Taking into account the permutation, the sparse Cholesky factor, $\mathbf{L}_\theta$, is defined to be the sparse, lower triangular, $q \times q$ matrix with positive diagonal elements satisfying

$$\mathbf{L}_\theta\mathbf{L}_\theta^\mathsf{T} = \mathbf{P}\left(\Lambda_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{Z}\Lambda_\theta + \mathbf{I}_q\right)\mathbf{P}^\mathsf{T}. \tag{5.22}$$

Note that we now require that the diagonal elements of Λ_θ be positive. Problems 5.1 and 5.2 indicate why we can require this. Because the diagonal elements of Λ_θ are positive, its determinant, $|\Lambda_\theta|$, which, for a triangular matrix such as Λ_θ, is simply the product of its diagonal elements, is also positive.

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 (5.22), 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 typically is 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 potentially non-zero at every non-zero location in 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 Davis [1996]. 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 details.

The pseudo-data representation of the PLS problem, (5.18), becomes

$$\tilde{\mathbf{u}} = \arg\min_{\mathbf{u}} \left\| \begin{bmatrix} \mathbf{y}_{\text{obs}} - \mathbf{X}\beta \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{Z}\Lambda_\theta \mathbf{P}^\mathsf{T} \\ \mathbf{P}^\mathsf{T} \end{bmatrix} \mathbf{P}\mathbf{u} \right\|^2 \tag{5.23}$$

and the system of linear equations satisfied by $\tilde{\mathbf{u}}$ is

$$\mathbf{L}_\theta \mathbf{L}_\theta^\mathsf{T} \mathbf{P}\tilde{\mathbf{u}} = \mathbf{P}\left(\Lambda_\theta^\mathsf{T} \mathbf{Z}^\mathsf{T} \mathbf{Z}\Lambda_\theta + \mathbf{I}_q\right) \mathbf{P}^\mathsf{T} \mathbf{P}\tilde{\mathbf{u}} = \mathbf{P}\Lambda_\theta^\mathsf{T} \mathbf{Z}^\mathsf{T} \left(\mathbf{y}_{\text{obs}} - \mathbf{X}\beta\right). \tag{5.24}$$

Obtaining the Cholesky factor, $\mathbf{L}_\theta$, may not seem to be great progress toward determining $\tilde{\mathbf{u}}$ because we still must solve (5.24) for $\tilde{\mathbf{u}}$. However, it is the key to the 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. Once we evaluate $\mathbf{L}_\theta$ it is straightforward to solve (5.24) for $\tilde{\mathbf{u}}$ because $\mathbf{L}_\theta$ is triangular.

In Sect. 5.6 we will describe the steps in determining this solution. First, though, we should show that the solution, $\tilde{\mathbf{u}}$, and the value of the objective at the solution, $r^2_{\theta,\beta}$, do allow us to evaluate the deviance.

5.4.2 The Value of the Deviance and Profiled Deviance

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

$$\|\mathbf{y}_{\text{obs}} - \mathbf{X}\beta - \mathbf{Z}\Lambda_\theta \mathbf{u}\|^2 + \|\mathbf{u}\|^2 = r^2_{\theta,\beta} + \|\mathbf{L}_\theta^\mathsf{T}(\mathbf{u} - \tilde{\mathbf{u}})\|^2 \tag{5.25}$$

which finally allows us to evaluate the likelihood. We plug the right hand side of (5.25) into the definition of $h(\mathbf{u})$ and apply the change of variable

$$\mathbf{z} = \frac{\mathbf{L}_\theta^\mathsf{T}(\mathbf{u} - \tilde{\mathbf{u}})}{\sigma}. \tag{5.26}$$

The determinant of the Jacobian of this transformation,

$$\left|\frac{d\mathbf{z}}{d\mathbf{u}}\right| = \left|\frac{\mathbf{L}_\theta^\mathsf{T}}{\sigma}\right| = \frac{|\mathbf{L}_\theta|}{\sigma^q} \tag{5.27}$$

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. That is, we will use the result

$$\int_{\mathbb{R}^q} \frac{e^{-\|\mathbf{z}\|^2/2}}{(2\pi)^{q/2}}\, d\mathbf{z} = 1. \tag{5.28}$$

Putting all these pieces together gives

$$\begin{aligned}
L(\theta,\beta,\sigma) &= \int_{\mathbb{R}^q} h(\mathbf{u})\, d\mathbf{u} \\
&= \int_{\mathbb{R}^q} \frac{1}{(2\pi\sigma^2)^{(n+q)/2}} \exp\left(-\frac{r^2_{\theta,\beta} + \|\mathbf{L}_\theta^\mathsf{T}(\mathbf{u}-\tilde{\mathbf{u}})\|^2}{2\sigma^2}\right) d\mathbf{u} \\
&= \frac{\exp\left(-\frac{r^2_{\theta,\beta}}{2\sigma^2}\right)}{(2\pi\sigma^2)^{n/2}} \int_{\mathbb{R}^q} \frac{1}{(2\pi)^{q/2}} \exp\left(-\frac{\|\mathbf{L}_\theta^\mathsf{T}(\mathbf{u}-\tilde{\mathbf{u}})\|^2}{2\sigma^2}\right) \frac{|\mathbf{L}_\theta|}{|\mathbf{L}_\theta|}\frac{d\mathbf{u}}{\sigma^q} \\
&= \frac{\exp\left(-\frac{r^2_{\theta,\beta}}{2\sigma^2}\right)}{(2\pi\sigma^2)^{n/2}|\mathbf{L}_\theta|} \int_{\mathbb{R}^q} \frac{e^{-\|\mathbf{z}\|^2/2}}{(2\pi)^{q/2}}\, d\mathbf{z} \\
&= \frac{\exp\left(-\frac{r^2_{\theta,\beta}}{2\sigma^2}\right)}{(2\pi\sigma^2)^{n/2}|\mathbf{L}_\theta|}.
\end{aligned} \tag{5.29}$$

The deviance can now be expressed as

$$d(\theta,\beta,\sigma|\mathbf{y}_{\text{obs}}) = -2\log\left(L(\theta,\beta,\sigma|\mathbf{y}_{\text{obs}})\right) = n\log(2\pi\sigma^2) + 2\log|\mathbf{L}_\theta| + \frac{r^2_{\beta,\theta}}{\sigma^2},$$

as stated in (1.6). The maximum likelihood estimates of the parameters are those that minimize this deviance.

Equation (1.6) 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 affects (1.6) only through $r^2_{\beta,\theta}$, and, for any value of θ, minimizing this expression with respect to β is just an extension of the penalized 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^2_θ be the PRSS at these minimizing values. If, in addition, we set $\widehat{\sigma^2}_\theta = r^2_\theta/n$, which is the value of σ^2 that minimizes the deviance for a given value of r^2_θ, then the *profiled deviance*, which is a function of θ only, becomes

$$\tilde{d}(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}}) = 2\log|\mathbf{L}_\theta| + n\left[1 + \log\left(\frac{2\pi r_\theta^2}{n}\right)\right]. \tag{5.30}$$

Numerical optimization (minimization) of $\tilde{d}(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}})$ with respect to $\boldsymbol{\theta}$ determines the MLE, $\widehat{\boldsymbol{\theta}}$. The MLEs for the other parameters, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\sigma}}$, are the corresponding conditional estimates evaluated at $\widehat{\boldsymbol{\theta}}$.

5.4.3 Determining r_θ^2 and $\widehat{\boldsymbol{\beta}}_\theta$

To determine $\tilde{\mathbf{u}}$ and $\widehat{\boldsymbol{\beta}}_\theta$ simultaneously we rearrange the terms in (5.23) as

$$\begin{bmatrix}\tilde{\mathbf{u}}\\ \widehat{\boldsymbol{\beta}}_\theta\end{bmatrix} = \arg\min_{\mathbf{u},\boldsymbol{\beta}} \left\| \begin{bmatrix}\mathbf{y}_{\text{obs}}\\ \mathbf{0}\end{bmatrix} - \begin{bmatrix}\mathbf{Z}\boldsymbol{\Lambda}_\theta\mathbf{P}^\mathsf{T} & \mathbf{X}\\ \mathbf{P}^\mathsf{T} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{P}\mathbf{u}\\ \boldsymbol{\beta}\end{bmatrix}\right\|^2. \tag{5.31}$$

The PLS values, $\tilde{\mathbf{u}}$ and $\widehat{\boldsymbol{\beta}}_\theta$, are the solutions to

$$\begin{bmatrix}\mathbf{P}\left(\boldsymbol{\Lambda}_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{Z}\boldsymbol{\Lambda}_\theta + \mathbf{I}_q\right)\mathbf{P}^\mathsf{T} & \mathbf{P}\boldsymbol{\Lambda}_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{X}\\ \mathbf{X}^\mathsf{T}\mathbf{Z}\boldsymbol{\Lambda}_\theta\mathbf{P}^\mathsf{T} & \mathbf{X}^\mathsf{T}\mathbf{X}\end{bmatrix}\begin{bmatrix}\mathbf{P}\tilde{\mathbf{u}}\\ \widehat{\boldsymbol{\beta}}_\theta\end{bmatrix} = \begin{bmatrix}\mathbf{P}\boldsymbol{\Lambda}_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{y}_{\text{obs}}\\ \mathbf{X}^\mathsf{T}\mathbf{y}_{\text{obs}}.\end{bmatrix} \tag{5.32}$$

To evaluate these solutions we decompose the system matrix as

$$\begin{bmatrix}\mathbf{P}\left(\boldsymbol{\Lambda}_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{Z}\boldsymbol{\Lambda}_\theta + \mathbf{I}_q\right)\mathbf{P}^\mathsf{T} & \mathbf{P}\boldsymbol{\Lambda}_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{X}\\ \mathbf{X}^\mathsf{T}\mathbf{Z}\boldsymbol{\Lambda}_\theta\mathbf{P}^\mathsf{T} & \mathbf{X}^\mathsf{T}\mathbf{X}\end{bmatrix} = \begin{bmatrix}\mathbf{L}_\theta & \mathbf{0}\\ \mathbf{R}_{ZX}^\mathsf{T} & \mathbf{R}_X^\mathsf{T}\end{bmatrix}\begin{bmatrix}\mathbf{L}_\theta^\mathsf{T} & \mathbf{R}_{ZX}\\ \mathbf{0} & \mathbf{R}_X\end{bmatrix} \tag{5.33}$$

where, as before, $\mathbf{L}_\theta$, the sparse Cholesky factor, is the sparse lower triangular $q \times q$ matrix satisfying (5.22). The other two matrices in (5.33): $\mathbf{R}_{ZX}$, which is a general $q \times p$ matrix, and $\mathbf{R}_X$, which is an upper triangular $p \times p$ matrix, satisfy

$$\mathbf{L}_\theta\mathbf{R}_{ZX} = \mathbf{P}\boldsymbol{\Lambda}_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{X} \tag{5.34}$$

and

$$\mathbf{R}_X^\mathsf{T}\mathbf{R}_X = \mathbf{X}^\mathsf{T}\mathbf{X} - \mathbf{R}_{ZX}^\mathsf{T}\mathbf{R}_{ZX} \tag{5.35}$$

Those familiar with standard ways of writing a Cholesky decomposition as either $\mathbf{L}\mathbf{L}^\mathsf{T}$ or $\mathbf{R}^\mathsf{T}\mathbf{R}$ ($\mathbf{L}$ is the factor as it appears on the left and $\mathbf{R}$ is as it appears on the right) will notice a notational inconsistency in (5.33). One Cholesky factor is defined as the lower triangular fractor on the left and the other is defined as the upper triangular factor on the right. It happens that in R the Cholesky factor of a dense positive-definite matrix is returned as the right factor, whereas the sparse Cholesky factor is returned as the left factor.

One other technical point that should be addressed is whether $\mathbf{X}^\mathsf{T}\mathbf{X} - \mathbf{R}_{ZX}^\mathsf{T}\mathbf{R}_{ZX}$ is positive definite. In theory, if $\mathbf{X}$ has full column rank, so that $\mathbf{X}^\mathsf{T}\mathbf{X}$ is positive definite, then the downdated matrix, $\mathbf{X}^\mathsf{T}\mathbf{X} - \mathbf{R}_{ZX}^\mathsf{T}\mathbf{R}_{ZX}$, must also be positive definite (see Prob. 5.4). In practice, the downdated matrix can

become computationally singular in ill-conditioned problems, in which case an error is reported.

The extended decomposition (5.33) not only provides for the evaluation of the profiled deviance function, $\tilde{d}(\theta)$, (5.30) but also allows us to define and evaluate the profiled REML criterion.

5.5 The REML Criterion

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,

$$\mathcal{Y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2\mathbf{I}_n), \tag{5.36}$$

in which we typically estimate σ^2 as

$$\widehat{\sigma^2_R} = \frac{\|\mathbf{y}_{\text{obs}} - \mathbf{X}\widehat{\beta}\|^2}{n-p} \tag{5.37}$$

even though the maximum likelihood estimate of σ^2 is

$$\widehat{\sigma^2_L} = \frac{\|\mathbf{y}_{\text{obs}} - \mathbf{X}\widehat{\beta}\|^2}{n}. \tag{5.38}$$

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, $\mathbf{y}_{\text{obs}} - \mathbf{X}\widehat{\beta}$, is an n-dimensional vector, the residual at $\widehat{\theta}$ satisfies p linearly independent constraints, $\mathbf{X}^\mathsf{T}(\mathbf{y}_{\text{obs}} - \mathbf{X}\widehat{\beta}) = \mathbf{0}$. That is, the residual at $\widehat{\theta}$ is the projection of the observed response vector, $\mathbf{y}_{\text{obs}}$, 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.*

Another argument often put forward for REML estimation is that $\widehat{\sigma^2_R}$ is an *unbiased* estimate of σ^2, in the sense that the expected value of the estimator is equal to the value of the parameter. However, determining the expected value of an estimator involves integrating with respect to the density of the estimator and we have seen that densities of estimators of variances will be skewed, often highly skewed. It is not clear why we should be interested in the expected value of a highly skewed estimator. If we were to transform to a more symmetric scale, such as the estimator of the standard deviation or the estimator of the logarithm of the standard deviation, the REML estimator

would no longer be unbiased. Furthermore, this property of unbiasedness of variance estimators does not generalize from the linear regression model to linear mixed models. This is all to say that the distinction between REML and ML estimates of variances and variance components is probably less important that many people believe.

Nevertheless it is worthwhile seeing how the computational techniques described in this chapter apply to the REML criterion because the REML parameter estimates $\widehat{\theta}_R$ and $\widehat{\sigma^2_R}$ for a linear mixed model have the property that they would specialize to $\widehat{\sigma^2_R}$ from (5.37) for a linear regression model, as seen in Sect. 1.3.2.

Although not usually derived in this way, the REML criterion (on the deviance scale) can be expressed as

$$d_R(\theta,\sigma|\mathbf{y}_{\text{obs}}) = -2\log\int_{\mathbb{R}^p} L(\theta,\beta,\sigma|\mathbf{y}_{\text{obs}})\,d\beta. \tag{5.39}$$

The REML estimates $\widehat{\theta}_R$ and $\widehat{\sigma^2_R}$ minimize $d_R(\theta,\sigma|\mathbf{y}_{\text{obs}})$.

To evaluate this integral we form an expansion, similar to (5.25), of $r^2_{\theta,\beta}$ about $\widehat{\beta}_\theta$

$$r^2_{\theta,\beta} = r^2_\theta + \|\mathbf{R}_X(\beta-\widehat{\beta}_\theta)\|^2. \tag{5.40}$$

In the same way that (5.25) was used to simplify the integral in (5.29), we can derive

$$\int_{\mathbb{R}^p}\frac{\exp\left(-\frac{r^2_{\theta,\beta}}{2\sigma^2}\right)}{(2\pi\sigma^2)^{n/2}|\mathbf{L}_\theta|}\,d\beta = \frac{\exp\left(-\frac{r^2_\theta}{2\sigma^2}\right)}{(2\pi\sigma^2)^{(n-p)/2}|\mathbf{L}_\theta||\mathbf{R}_X|} \tag{5.41}$$

corresponding to a REML criterion on the deviance scale of

$$d_R(\theta,\sigma|\mathbf{y}_{\text{obs}}) = (n-p)\log(2\pi\sigma^2) + 2\log\left(|\mathbf{L}_\theta||\mathbf{R}_X|\right) + \frac{r^2_\theta}{\sigma^2}. \tag{5.42}$$

Plugging in the conditional REML estimate, $\widehat{\sigma^2}_R = r^2_\theta/(n-p)$, provides the profiled REML criterion

$$\tilde{d}_R(\theta|\mathbf{y}_{\text{obs}}) = 2\log\left(|\mathbf{L}_\theta||\mathbf{R}_X|\right) + (n-p)\left[1+\log\left(\frac{2\pi r^2_\theta}{n-p}\right)\right] \tag{5.43}$$

The REML estimate of θ is

$$\widehat{\theta}_R = \arg\min_\theta \tilde{d}_R(\theta|\mathbf{y}_{\text{obs}}), \tag{5.44}$$

and the REML estimate of σ^2 is the conditional REML estimate of σ^2 at $\widehat{\theta}_R$,

$$\widehat{\sigma^2_R} = r^2_{\widehat{\theta}_R}/(n-p). \tag{5.45}$$

It is not entirely clear how one would define a "REML estimate" of β because the REML criterion, $d_R(\theta, \sigma|\mathbf{y})$, defined in (5.42), does not depend on β. However, it is customary (and not unreasonable) to use $\widehat{\beta}_R = \widehat{\beta}_{\widehat{\theta}_R}$ as the REML estimate of β.

5.6 Step-by-step Evaluation of the Profiled Deviance

As we have seen, an object returned by `lmer` contains an environment, accessed with the `env` extractor. This environment contains several matrices and vectors that are used in the evaluation of the profiled deviance. In this section we use these matrices and vectors from one of our examples to explicitly trace the steps in evaluating the profiled deviance. This level of detail is provided for those whose style of learning is more of a "hands on" style and for those who may want to program modifications of this approach.

Consider our model `fm8`, fit as

```
> fm8 <- lmer(Reaction ~ 1 + Days + (1 + Days|Subject), sleepstudy,
+             REML = 0, verbose = TRUE)

  0:     1784.6423:  1.00000  0.00000  1.00000
  1:     1774.2946:  1.00042 -0.00836471 0.725280
  2:     1754.3212: 0.998969 -0.0239943 0.175808
  3:     1752.1500: 0.998284 0.00682229 0.243192
...
```

The environment of the model contains the converged parameter vector, θ (`theta`), the relative covariance factor, Λ_θ (`Lambda`), the sparse Cholesky factor, $\mathbf{L}_\theta$ (`L`), the matrices $\mathbf{R}_{ZX}$ (`RZX`) and $\mathbf{R}_X$ (`RX`), the conditional mode, $\tilde{\mathbf{u}}$ (`u`), and the conditional estimate, $\widehat{\beta}_\theta$ (`fixef`). The permutation represented by $\mathbf{P}$ is contained in the sparse Cholesky representation, `L`.

Although the model matrices, $\mathbf{X}$ (`X`) and $\mathbf{Z}^\mathsf{T}$ (`Zt`), and the response vector, $\mathbf{y}_{\text{obs}}$ (`y`), are available in the environment, many of the products that involve only these fixed values are precomputed and stored separately under the names `XtX` ($\mathbf{X}^\mathsf{T}\mathbf{X}$), `Xty`, `ZtX` and `Zty`.

To provide easy access to the objects in the environment of `fm8` we attach it to the search path.

```
> attach(env(fm8))
```

Please note that this is done here for illustration only. The practice of attaching a list or a data frame or, less commonly, an environment in an R session is overused, somewhat dangerous (because of the potential of forgetting to detach it later) and discouraged. The preferred practice is to use the `with` function to gain access by name to components of such composite objects. For this section of code, however, using `with` or `within` would quickly become very tedious and we use `attach` instead.

To update the matrix Λ_θ to a new value of θ we need to know which of the non-zeros in Λ are updated from which elements of θ. Recall that the dimension of θ is small (3, in this case) but Λ is potentially large (18×18 with 54 non-zeros). The environment contains an integer vector `Lind` that maps the elements of `theta` to the non-zeros in `Lambda`.

Suppose we wish to recreate the evaluation of the profiled deviance at the initial value of $\theta = (1, 0, 1)$. We begin by updating Λ_θ and forming the product $\mathbf{U}^\mathsf{T} = \Lambda_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}$

```
> str(Lambda)

Formal class 'dgCMatrix' [package "Matrix"] with 6 slots
  ..@ i       : int [1:54] 0 1 1 2 3 3 4 5 5 6 ...
  ..@ p       : int [1:37] 0 2 3 5 6 8 9 11 12 14 ...
  ..@ Dim     : int [1:2] 36 36
  ..@ Dimnames:List of 2
  .. ..$ : NULL
  .. ..$ : NULL
  ..@ x       : num [1:54] 0.9292 0.0182 0.2226 0.9292 0.0182 ...
  ..@ factors : list()

> str(Lind)

 int [1:54] 1 2 3 1 2 3 1 2 3 1 ...

> Lambda@x[] <- c(1,0,1)[Lind]
> str(Lambda@x)

 num [1:54] 1 0 1 1 0 1 1 0 1 1 ...

> Ut <- crossprod(Lambda, Zt)
```

The Cholesky factor object, `L`, can be updated from `Ut` without forming $\mathbf{U}^\mathsf{T}\mathbf{U} + \mathbf{I}$ explicitly. The optional argument `mult` to the `update` method specifies a multiple of the identity to be added to $\mathbf{U}^\mathsf{T}\mathbf{U}$

```
> L <- update(L, Ut, mult = 1)
```

Then we evaluate `RZX` and `RX` according to (5.34) and (5.35)

```
> RZX <- solve(L, solve(L, crossprod(Lambda, ZtX), sys = "P"), sys = "L")
> RX <- chol(XtX - crossprod(RZX))
```

Solving (5.32) for $\tilde{\mathbf{u}}$ and $\widehat{\beta}_\theta$ is done in stages. Writing $\mathbf{c}_u$ and $\mathbf{c}_\beta$ for the intermediate results that satisfy

$$\begin{bmatrix} \mathbf{L}_\theta & \mathbf{0} \\ \mathbf{R}_{ZX}^\mathsf{T} & \mathbf{R}_X^\mathsf{T} \end{bmatrix} \begin{bmatrix} \mathbf{c}_u \\ \mathbf{c}_\beta \end{bmatrix} = \begin{bmatrix} \mathbf{P}\Lambda_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{y}_{\text{obs}} \\ \mathbf{X}^\mathsf{T}\mathbf{y}_{\text{obs}}. \end{bmatrix} \tag{5.46}$$

we evaluate

```
> cu <- solve(L, solve(L, crossprod(Lambda, Zty), sys = "P"), sys = "L")
> cbeta <- solve(t(RX), Xty - crossprod(RZX, cu))
```

The next set of equations to solve is

$$\begin{bmatrix} \mathbf{L}_\theta^\top & \mathbf{R}_{ZX} \\ \mathbf{0} & \mathbf{R}_X \end{bmatrix} \begin{bmatrix} \mathbf{P}\tilde{\mathbf{u}} \\ \widehat{\beta}_\theta \end{bmatrix} = \begin{bmatrix} \mathbf{c}_U \\ \mathbf{c}_\beta \end{bmatrix} \tag{5.47}$$

```
> fixef <- as.vector(solve(RX, cbeta))
> u <- solve(L, solve(L, cu - RZX %*% fixef, sys = "Lt"), sys = "Pt")
```

We can now create the conditional mean, `mu`, the penalized residual sum of squares, `prss`, the logarithm of the square of the determinant of $\mathbf{L}$, `ldL2`, and the profiled deviance, which, fortuitously, equals the value shown earlier.

```
> mu <- gamma <- as.vector(crossprod(Ut, u) + X %*% fixef)
> prss <- sum(c(y - mu, as.vector(u))^2)
> ldL2 <- 2 * as.vector(determinant(L)$mod)
> (deviance <- ldL2 + nobs * (1 + log(2 * pi * prss/nobs)))
```

```
[1] 1784.642
```

The last step is detach the environment of `fm8` from the search list

```
> detach()
```

to avoid later name clashes.

In terms of the calculations performed, these steps describe exactly the evaluation of the profiled deviance in `lmer`. The actual function for evaluating the deviance, accessible as `fm8@setPars`, is a slightly modified version of what is shown above. However, the modifications are only to avoid creating copies of potentially large objects and to allow for cases where the model matrix, $\mathbf{X}$, is sparse. In practice, unless the optional argument `compDev = FALSE` is given, the profiled deviance is evaluated in compiled code, providing a speed boost, but the R code can be used if desired. This allows for checking the results from the compiled code and can also be used as a template for extending the computational methods to other types of models.

5.7 Generalizing to Other Forms of Mixed Models

In later chapters we cover the theory and practice of generalized linear mixed models (GLMMs), nonlinear mixed models (NLMMs) and generalized nonlinear mixed models (GNLMMs). Because quite a bit of the theoretical and computational methodology covered in this chapter extends to those models we will cover the common aspects here.

5.7.1 Descriptions of the Model Forms

We apply the name "generalized" to models in which the conditional distribution, $(\mathscr{Y}|\mathscr{U}=\mathbf{u})$, is not required to be Gaussian but does preserve some of the properties of the spherical Gaussian conditional distribution

$$(\mathscr{Y}|\mathscr{U}=\mathbf{u})\sim\mathscr{N}(\mathbf{Z}\Lambda_\theta\mathbf{u}+\mathbf{X}\beta,\sigma^2\mathbf{I}_n)$$

from the linear mixed model. In particular, the components of $\mathscr{Y}$ are *conditionally independent*, given $\mathscr{U}=\mathbf{u}$. Furthermore, $\mathbf{u}$ affects the distribution only through the conditional mean, which we will continue to write as μ, and it affects the conditional mean only through the linear predictor, $\gamma=\mathbf{Z}\Lambda_\theta\mathbf{u}+\mathbf{X}\beta$.

Typically we do not have $\mu=\gamma$, however. The elements of the linear predictor, γ, can be positive or negative or zero. Theoretically they can take on any value between $-\infty$ and ∞. But many distributional forms used in GLMMs put constraints on the value of the mean. For example, the mean of a Bernoulli random variable, modeling a binary response, must be in the range $0<\mu<1$ and the mean of a Poisson random variable, modeling a count, must be positive. To achieve these constraints we write the conditional mean, μ as a transformation of the unbounded predictor, written η. For historical, and some theoretical, reasons the inverse of this transformation is called the *link function*, written

$$\eta=\mathbf{g}(\mu), \tag{5.48}$$

and the transformation we want is called the *inverse link*, written $\mathbf{g}^{-1}$.

Both $\mathbf{g}$ and $\mathbf{g}^{-1}$ are determined by scalar functions, g and g^{-1}, respectively, applied to the individual components of the vector argument. That is, η must be n-dimensional and the vector-valued function $\mu=\mathbf{g}^{-1}(\eta)$ is defined by the component functions $\mu_i=g^{-1}(\eta_i)$, $i=1,\ldots,n$. Among other things, this means that the Jacobian matrix of the inverse link, $\frac{d\mu}{d\eta}$, will be diagonal.

Because the link function, $\mathbf{g}$, and the inverse link, $\mathbf{g}^{-1}$, are nonlinear functions (there would be no purpose in using a linear link function) many people use the terms "generalized linear mixed model" and "nonlinear mixed model" interchangably. We reserve the term "nonlinear mixed model" for the type of models used, for example, in pharmacokinetics and pharmacodynamics, where the conditional distribution is a spherical multivariate Gaussian

$$(\mathscr{Y}|\mathscr{U}=\mathbf{u})\sim\mathscr{N}(\mu,\sigma^2\mathbf{I}_n) \tag{5.49}$$

but μ depends nonlinearly on γ. For NLMMs the length of the linear predictor, γ, is a multiple, ns, of n, the length of μ.

Like the map from η to μ, the map from γ to μ has a "diagonal" property, which we now describe. If we use γ to fill the columns of an $n\times s$ matrix, Γ, then μ_i depends only on the ith row of Γ. In fact, μ_i is determined by a

nonlinear model function, f, applied to the i row of Γ. Writing $\mu = \mathbf{f}(\gamma)$ based on the component function f, we see that the Jacobian of $\mathbf{f}$, $\frac{d\mu}{d\gamma}$, will be the vertical concatenation of s diagonal $n \times n$ matrices.

Because we will allow for generalized nonlinear mixed models (GNLMMs), in which the mapping from γ to μ has the form

$$\gamma \rightarrow \eta \rightarrow \mu, \tag{5.50}$$

we will use (5.50) in our definitions.

5.7.2 Determining the Conditional Mode, $\tilde{\mathbf{u}}$

For all these types of mixed models, the conditional distribution, $(\mathscr{U}|\mathscr{Y} = \mathbf{y}_{\text{obs}})$ is a continuous distribution for which we can determine the unscaled conditional density, $h(\mathbf{u})$. As for linear mixed models, we define the conditional mode, $\tilde{\mathbf{u}}$ as the value that maximizes the unscaled condtional density.

Determining the conditional mode, $\tilde{\mathbf{u}}$, in a nonlinear mixed model is a penalized nonlinear least squares (PNLS) problem

$$\tilde{\mathbf{u}} = \arg\min_{\mathbf{u}} \|\mathbf{y}_{\text{obs}} - \mu\|^2 + \|\mathbf{u}\|^2 \tag{5.51}$$

which we solve by adapting the iterative techniques, such as the Gauss-Newton method [Bates and Watts, 1988, Sect. 2.2.1], used for nonlinear least squares. Starting at an initial value, $\mathbf{u}^{(0)}$, (the bracketed superscript denotes the iteration number) with conditional mean, $\mu^{(0)}$, we determine an increment $\delta^{(1)}$ by solving the penalized linear least squares problem,

$$\delta^{(1)} = \arg\min_{\delta} \left\| \begin{bmatrix} \mathbf{y}_{\text{obs}} - \mu^{(0)} \\ \mathbf{0} - \mathbf{u}^{(0)} \end{bmatrix} - \begin{bmatrix} \mathbf{U}^{(0)} \\ \mathbf{I}_q \end{bmatrix} \delta \right\|^2 \tag{5.52}$$

where

$$\mathbf{U}^{(0)} = \left.\frac{d\mu}{d\mathbf{u}}\right|_{\mathbf{u}^{(0)}}. \tag{5.53}$$

Naturally, we use the sparse Cholesky decomposition, $\mathbf{L}_\theta^{(0)}$, satisfying

$$\mathbf{L}_\theta^{(0)} \left(\mathbf{L}_\theta^{(0)}\right) = \mathbf{P}\left[\left(\mathbf{U}^{(0)}\right)^{\mathsf{T}} \mathbf{U}^{(0)} + \mathbf{I}_q\right] \mathbf{P}^{\mathsf{T}} \tag{5.54}$$

to determine this increment. The next iteration begins at

$$\mathbf{u}^{(1)} = \mathbf{u}^{(0)} + k\delta^{(1)} \tag{5.55}$$

where k is the step factor chosen, perhaps by step-halving [Bates and Watts, 1988, Sect. 2.2.1], to ensure that the penalized residual sum of squares decreases at each iteration. Convergence is declared when the orthogonality convergence criterion [Bates and Watts, 1988, Sect. 2.2.3] is below some prespecified tolerance.

The *Laplace approximation* to the deviance is

$$d(\theta,\beta,\sigma|\mathbf{y}_{\text{obs}}) \approx= n\log(2\pi\sigma^2)+2\log|\mathbf{L}_{\theta,\beta}|+\frac{r^2_{\theta,\beta}}{\sigma^2}, \tag{5.56}$$

where the Cholesky factor, $\mathbf{L}_{\theta,\beta}$, and the penalized residual sum of squares, $r^2_{\theta,\beta}$, are both evaluated at the conditional mode, $\tilde{\mathbf{u}}$. The Cholesky factor depends on θ, β and $\mathbf{u}$ for these models but typically the dependence on β and $\mathbf{u}$ is weak.

5.8 Chapter Summary

The definitions and the computational results for maximum likelihood estimation of the parameters in linear mixed models were summarized in Sect. 1.4.1. A key computation is evaluation of the sparse Cholesky factor, Λ_θ, satisfying eqn. 5.22,

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

where $\mathbf{P}$ represents the fill-reducing permutation determined during the symbolic phase of the sparse Cholesky decomposition.

An extended decomposition (eqn. 5.33) provides the $q\times p$ matrix $\mathbf{R}_{ZX}$ and the $p\times p$ upper triangular $\mathbf{R}_X$ that are used to determine the conditional mode $\tilde{\mathbf{u}}_\theta$, the conditional estimate $\widehat{\beta}_\theta$ and the minimum penalized residual sum of squares, r^2_θ from which the profiled deviance

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

or the profile REML criterion

$$\tilde{d}_R(\theta|\mathbf{y}_{\text{obs}}) = 2\log\left(|\mathbf{L}_\theta||\mathbf{R}_X|\right)+(n-p)\left[1+\log\left(\frac{2\pi r^2_\theta}{n-p}\right)\right]$$

can be evaluated and optimized (minimized) with respect to θ.

Exercises

Unlike the exercises in other chapters, these exercises establish theoretical results, which do not always apply exactly to the computational results.

5.1. Show that the matrix $\mathbf{A}_\theta = \mathbf{P}\boldsymbol{\Lambda}_\theta^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{Z}\boldsymbol{\Lambda}_\theta\mathbf{P}^\mathsf{T} + \mathbf{I}_q$ is positive definite. That is, $\mathbf{b}^\mathsf{T}\mathbf{A}\mathbf{b} > 0, \forall \mathbf{b} \neq \mathbf{0}$.

5.2. (a) Show that $\boldsymbol{\Lambda}_\theta$ can be defined to have non-negative diagonal elements. (Hint: Show that the product $\boldsymbol{\Lambda}_\theta\mathbf{D}$ where $\mathbf{D}$ is a diagonal matrix with diagonal elements of ± 1 is also a Cholesky factor. Thus the signs of the diagonal elements can be chosen however we want.)

(b) Use the result of Prob. 5.1 to show that the diagonal elements of $\boldsymbol{\Lambda}_\theta$ must be non-zero. (Hint: Suppose that the first zero on the diagonal of $\boldsymbol{\Lambda}_\theta$ is in the ith position. Show that there is a solution $\mathbf{x}$ to $\boldsymbol{\Lambda}_\theta^\mathsf{T}\mathbf{x} = \mathbf{0}$ with $x_i = 1$ and $x_j = 0$, $j = i+1, \ldots, q$ and that this $\mathbf{x}$ contradicts the positive definite condition.)

5.3. Show that if $\mathbf{X}$ has full column rank, which means that there does not exist a $\boldsymbol{\beta} \neq \mathbf{0}$ for which $\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$, then $\mathbf{X}^\mathsf{T}\mathbf{X}$ is positive definite.

5.4. Show that if $\mathbf{X}$ has full column rank then

$$\begin{bmatrix} \mathbf{Z}\boldsymbol{\Lambda}_\theta\mathbf{P}^\mathsf{T} & \mathbf{X} \\ \mathbf{P}^\mathsf{T} & \mathbf{0} \end{bmatrix}$$

also must have full column rank. (Hint: First show that $\mathbf{u}$ must be zero in any vector $\begin{bmatrix} \mathbf{u} \\ \boldsymbol{\beta} \end{bmatrix}$ satisfying

$$\begin{bmatrix} \mathbf{Z}\boldsymbol{\Lambda}_\theta\mathbf{P}^\mathsf{T} & \mathbf{X} \\ \mathbf{P}^\mathsf{T} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\beta} \end{bmatrix} = \mathbf{0}.$$

Use this result and (5.33) to show that

$$\begin{bmatrix} \mathbf{L}_\theta & \mathbf{0} \\ \mathbf{R}_{ZX}^\mathsf{T} & \mathbf{R}_X^\mathsf{T} \end{bmatrix} \begin{bmatrix} \mathbf{L}_\theta^\mathsf{T} & \mathbf{R}_{ZX} \\ \mathbf{0} & \mathbf{R}_X \end{bmatrix}$$

is positive definite and, hence, $\mathbf{R}_X$ is non-singular.)