Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

blme: Profiled Maximum A Posteriori Estimation of Bayesian Linear Mixed Effects Models in R

Vincent Dorie
New York University

Abstract

The popular R package **lme4** allows the fast fitting of linear and generalized linear mixed effect models through the use of a profiled likelihood function. This article details Bayesian, or penalized-likelihood, extensions that enable a wider class of models to be fit while preserving computational efficiency. The package **blme**, also for R, implements a wide variety of these models while inheriting functionality from **lme4**. Projected uses include the incorporation of substantive prior information, regularization of fixed effects, meta analysis, and the use of weakly informative priors to obtain non-degenerate random effect covariance matrix estimates.

Keywords: Bayesian, penalized likelihood, mixed effects, **blme**, **lme4**, R.

1. Introduction

History, references for lmm, glmms. AKA.
Proof of popularity? Popularity of lme4?
lmms lme4 profiles; analysis of profiling technique shows what additional models can fit
map, penalized likelihood; meta analysis, ridge regression; boundary estimates
competing packages
things you can do, not whether or not you should do them

2. Profiled likelihoods and linear mixed models

In the next few sections, we derive Bayesian extensions / likelihood penalties that can be applied to the linear mixed model fit by **lme4** with little to no additional computation cost. To do so, we first trace through the calculations that **lme4** uses to profile the likelihood and

avoid costly numeric optimization techniques, after which the modifications are introduced. A few relevant concepts, such as common scale parameters and REML estimation are introduced as well. When comparing two equations where one has been altered to incorporate a prior, the key differences are highlighted in blue text.

We also note that this section applies only to *linear* mixed models. It is the approach of profiling that makes adding priors difficult, and as generalized linear mixed models in **lme4** are fit without profiling, priors are considerably easier to implement. They are briefly discussed in section 4.

2.1. lme4 linear mixed model specification

In concise, matrix notation the linear mixed effect model fit by **lme4** can be written as

$$\begin{aligned} y \mid b &\sim \mathcal{N}(X\beta + Zb, \sigma^2 I_n), \\ b &\sim \mathcal{N}(0, \sigma^2 \Sigma), \end{aligned} \tag{1}$$

where X and Z are matrices of suitable dimension, β are the fixed effects, b are the random effects, σ^2 is the residual variance or "common scale" parameter, and Σ is the covariance of the random effects. Furthermore, Z and Σ are sparse, as Z serves to select the specific random effects that are applicable to an observation and the random effects are independent between groups and grouping factors. The structure of Σ is irrelevant to maximum likelihood so we ignore it for the moment, although it is explored in-depth in section 3.4 when random effect covariance priors are considered. As b is latent, integrating it out yields the marginal model wherein

$$y \sim \mathcal{N}(X\beta, \sigma^2 I_n + \sigma^2 Z\Sigma Z^\top), \tag{2}$$

subject to the constraint that Σ is positive semi-definite. The parameters are β, σ^2, and Σ.

A reader familiar with generalized least-squares problems might notice that, for a given value of Σ, the maximum likelihood estimates of β and σ^2 are easy to obtain. That is indeed the case, but rather than directly optimize the associated likelihood and repeatedly invert a matrix of dimension equal to the sample size, the authors of **lme4** employ a series of profiling steps that exploit the sparsity of the problem. We retrace their solution here, as it has implications for what priors can be imposed without adding substantial computation.

2.2. Common scale

Here we take a brief aside to discuss the implications of modeling the random effects on the same scale as the observations, that is the appearance of σ^2 on the second line of equation 1. As will soon be shown, the choice greatly simplifies optimization. However, for the purposes of the extending the model it essentially requires that priors also be placed on this common scale if the efficiency gains of **lme4** are to be preserved.

In the absence of substantive prior knowledge, there do not seem to be any strong reasons to want to influence the model in ways that aren't modulo the residual variance. However, when such knowledge does exist incorporating it sometimes comes at the cost of mathematical convenience and profiling must be abandoned. Exceptions will be noted as they occur.

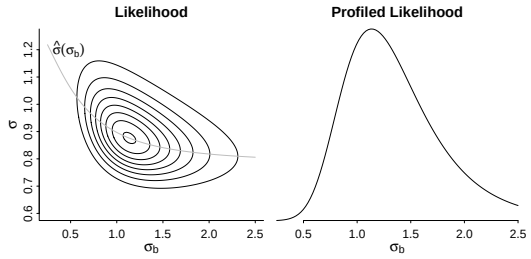

Figure 1: Illustration of a profiled likelihood. The left graph shows the likelihood for a simple linear model mixed model with a varying intercept as a function of σ and $\sqrt{\Sigma}$, which we abbreviate as σ_b (β has already been maximized). The gray line corresponds to the maximum in σ as a function of σ_b. Following the contour along this line produces the profiled likelihood of σ_b on the right. The profiled likelihood goes through the joint mode, so that maximizing it is sufficient to maximize the likelihood.

2.3. Restricted/residual maximum likelihood

As a final preliminary, we consider restricted maximum likelihood estimation, abbreviated REML. REML is commonly used as to unbias the estimate of the conditional variance. In one sense, it takes into account the cost of estimating the fixed effects, although for our purposes it is sufficient to know that it also arises from integrating out the fixed effects from the likelihood using a flat prior. Since this assumption is rarely made explicit, we refer to REML as an estimation procedure and not a model. Consequently, in REML estimation an objective function/criterion is maximized instead of a likelihood. This function we label 'q', and refer to proper probability densities by 'p'.

2.4. Definition of profiled likelihood

A profiled likelihood is a function over several parameters, some of which have been optimized analytically. Once a maximizer for one or more parameters has been derived, these estimators can be plugged back into the likelihood and the resulting equation optimized instead.

Figure 1 demonstrates this approach for a simple linear mixed effects model with only a varying intercept. The likelihood is a function of the three parameters β, σ^2, and Σ - all of which in this case are scalars. The estimator of β that maximizes the likelihood is a function of the two variance parameters and has an explicit solution. Plugging this into the likelihood yields a first-stage profiled function. Similarly, the maximizer of this function with respect to σ^2 can be obtained as a function of just Σ. These two steps produce a profiled likelihood that is a function only of the variance of the random effects, here reducing the problem from three parameters to just one.

2.5. Profiled likelihood derivation

One of the key features of **lme4** for linear mixed models is that it performs profiled optimization by analytically maximizing the likelihood in first the fixed effects and subsequently the residual variance. Only the covariance of the random effect requires numeric techniques. In this section, we recreate the derivation of the profiled likelihood so that it is possible to demonstrate which Bayesian extensions fit in this two-step profiling scheme.

We start with the joint density of the response and random effects. The random effects must be integrated from this equation to obtain the likelihood, but we first manipulate it so as to simplify subsequent optimization. To be precise, denote N as the length of y, or the total number of observations. Let Q be the total number of random effects and P the number of fixed effects. Then,

$$p(y, b; \beta, \sigma^2, \Sigma) = (2\pi\sigma^2)^{-(N+Q)/2}|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}\left[\|y - X\beta - Zb\|^2 + b^\top \Sigma^{-1} b\right]\right\}.$$

As we intend to integrate out b, we can make the change of variables to spherical random effects. If $\Lambda\Lambda^\top = \Sigma$ is a left-Cholesky factorization of Σ, then for $b = \Lambda u, u \sim \mathcal{N}(0, I_Q)$ we have

$$\begin{aligned} p(y, u; \beta, \sigma^2, \Lambda) &= (2\pi\sigma^2)^{-(N+Q)/2} \exp\left\{-\frac{1}{2\sigma^2}\left[\|y - X\beta - Z\Lambda u\|^2 + \|u\|^2\right]\right\}, \\ &= (2\pi\sigma^2)^{-(N+Q)/2} \exp\left\{-\frac{1}{2\sigma^2}\left\|\begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} Z\Lambda & X \\ I_Q & 0 \end{bmatrix}\begin{bmatrix} u \\ \beta \end{bmatrix}\right\|^2\right\}. \end{aligned} \tag{3}$$

At this point it is possible to identify the maximum likelihood estimate of β. For the moment conceptualizing β as a random variable, the preceeding equation has a quadratic form in the exponential term for the vector (u, β). In turn, this defines the two as jointly-normal and implies that the marginal mode for β is the same as the joint mode. As we now detail, the MLE is obtained by finding the mode of this joint distribution in u and β, "conditioning" on β, and integrating out $u \mid \beta$.

Denote $\tilde{u}$ and $\tilde{\beta}$ as the modes of this joint distribution. They are obtained in a nuanced fashion that exploits sparse matrix decompositions and is reproduced for completeness in appendix 8.1. The decomposition that is used is given by the blocks of

$$\begin{aligned} \begin{bmatrix} L_Z & 0 \\ L_{ZX} & L_X \end{bmatrix}\begin{bmatrix} L_Z^\top & L_{ZX}^\top \\ 0 & L_X^\top \end{bmatrix} &= \begin{bmatrix} \Lambda^\top Z^\top & I_Q \\ X^\top & 0 \end{bmatrix}\begin{bmatrix} Z\Lambda & X \\ I_Q & 0 \end{bmatrix}, \\ &= \begin{bmatrix} \Lambda^\top Z^\top Z\Lambda + I_Q & \Lambda^\top Z^\top X \\ X^\top Z\Lambda & X^\top X \end{bmatrix}. \end{aligned}$$

Specifically,

$$\begin{aligned} L_Z L_Z^\top &= \Lambda^\top Z^\top Z\Lambda + I_Q, \\ L_{ZX} &= X^\top Z\Lambda L_Z^{-\top}, \\ L_X L_X^\top &= X^\top X - L_{ZX} L_{ZX}^\top. \end{aligned} \tag{4}$$

As Z is sparse and Λ is block-diagonal, L_Z can be computed and stored efficiently. While we typically omit the dependence, it is important to note that these matrices depend on Λ.

Utilizing this decomposition and the joint modes, we can rewrite the joint distribution by completing the square:

$$p(y,u;\beta,\sigma^2,\Lambda) = (2\pi\sigma^2)^{-(N+Q)/2}\exp\left\{-\frac{1}{2\sigma^2}\left(\begin{bmatrix}u-\tilde{u}\\ \beta-\tilde{\beta}\end{bmatrix}^\top \begin{bmatrix}L_Z & 0\\ L_{ZX} & L_X\end{bmatrix}\begin{bmatrix}L_Z^\top & L_{ZX}^\top\\ 0 & L_X^\top\end{bmatrix}\begin{bmatrix}u-\tilde{u}\\ \beta-\tilde{\beta}\end{bmatrix} + \left\|\begin{bmatrix}y\\ 0\end{bmatrix} - \begin{bmatrix}Z\Lambda & X\\ I_Q & 0\end{bmatrix}\begin{bmatrix}\tilde{u}\\ \tilde{\beta}\end{bmatrix}\right\|^2\right)\right\}.$$

Conditioning on β simply rotates its dependence with u into its mean, so that writing $\mu_{u|\beta} = \tilde{u} - L_Z^{-\top}L_{ZX}^\top(\beta-\tilde{\beta})$ we have:

$$p(y,u;\beta,\sigma^2,\Lambda) = (2\pi\sigma^2)^{-(N+Q)/2}\exp\left\{-\frac{1}{2\sigma^2}\left(\begin{bmatrix}u-\mu_{u|\beta}\\ \beta-\tilde{\beta}\end{bmatrix}^\top \begin{bmatrix}L_Z & 0\\ 0 & L_X\end{bmatrix}\begin{bmatrix}L_Z^\top & 0\\ 0 & L_X^\top\end{bmatrix}\begin{bmatrix}u-\mu_{u|\beta}\\ \beta-\tilde{\beta}\end{bmatrix} + \left\|\begin{bmatrix}y\\ 0\end{bmatrix} - \begin{bmatrix}Z\Lambda & X\\ I_Q & 0\end{bmatrix}\begin{bmatrix}\tilde{u}\\ \tilde{\beta}\end{bmatrix}\right\|^2\right)\right\}.$$

At this point, it is trivial to integrate out u as it has the density of Gaussian random variable which we can factor in the form of $p(u \mid \beta)p(\beta)$. Performing this integration, we obtain the rearranged marginal likelihood

$$p(y;\beta,\sigma^2,\Lambda) = (2\pi\sigma^2)^{-N/2}|L_Z|^{-1}\exp\left\{-\frac{1}{2\sigma^2}\left[(\beta-\tilde{\beta})^\top L_X L_X^\top(\beta-\tilde{\beta}) + \left\|\begin{bmatrix}y\\ 0\end{bmatrix} - \begin{bmatrix}Z\Lambda & X\\ I_Q & 0\end{bmatrix}\begin{bmatrix}\tilde{u}\\ \tilde{\beta}\end{bmatrix}\right\|^2\right]\right\}. \quad (5)$$

Having the likelihood in this form makes the profiling steps straightforward. The maximum likelihood estimator of β is $\hat{\beta} = \tilde{\beta}$, which is also the joint mode. Plugging this back in gives us the first stage profiled equation

$$p(y;\hat{\beta},\sigma^2,\Lambda) = (2\pi\sigma^2)^{-N/2}|L_Z|^{-1}\exp\left\{-\frac{1}{2\sigma^2}\left\|\begin{bmatrix}y\\ 0\end{bmatrix} - \begin{bmatrix}Z\Lambda & X\\ I_Q & 0\end{bmatrix}\begin{bmatrix}\tilde{u}\\ \tilde{\beta}\end{bmatrix}\right\|^2\right\}.$$

From this, maximizing in σ^2 yields the MLE $\hat{\sigma}^2 = \frac{1}{N}\left[\|y - X\hat{\beta} - Z\Lambda\tilde{u}\|^2 + \|\tilde{u}\|^2\right]$. Utilizing this in turn produces the fully profiled likelihood:

$$p(y;\hat{\beta},\hat{\sigma}^2,\Lambda) = \left(2\pi\hat{\sigma}^2(\Lambda)\right)^{-N/2}|L_Z(\Lambda)|^{-1}e^{-N/2}. \quad (6)$$

REML estimation proceeds by integrating β from the likelihood as a Gaussian random variable rather than by maximization. The REML objective function is

$$q(y;\sigma^2,\Lambda) = (2\pi\sigma^2)^{-(N-P)/2}|L_Z|^{-1}|L_X|^{-1}\exp\left\{-\frac{1}{2\sigma^2}\left\|\begin{bmatrix}y\\ 0\end{bmatrix} - \begin{bmatrix}Z\Lambda & X\\ I_Q & 0\end{bmatrix}\begin{bmatrix}\tilde{u}\\ \tilde{\beta}\end{bmatrix}\right\|^2\right\}. \quad (7)$$

The REML estimate of σ^2 is $\hat{\sigma}^2_{\text{RE}} = \frac{1}{N-P}\left[\|y - X\tilde{\beta} - Z\Lambda\tilde{u}\|^2 + \|\tilde{u}\|^2\right]$, while the profiled REML objective function is

$$q(y; \hat{\sigma}^2, \Lambda) = \left(2\pi\hat{\sigma}^2_{\text{RE}}(\Lambda)\right)^{-(N-P)/2} |L_Z(\Lambda)|^{-1}|L_X(\Lambda)|^{-1}e^{-(N-P)/2}. \tag{8}$$

3. Profiled posteriors for linear mixed models

In this section we describe Bayesian extensions to linear mixed models which can be profiled in fashion similar to the two-step approached used for the likelihood. Arbitrary priors can be applied and the posterior mode found instead, but doing so may dramatically increase the number of parameters that require numeric optimization.

3.1. Preliminaries

As priors are successively applied to model components it becomes important to clearly define the quantity that is being estimated. For example, when placing a prior over the fixed effects (β) but not the other parameters, the traditional Bayesian estimand is the posterior distribution $p(u, \beta|y; \sigma^2, \Sigma)$ while the variance components (σ^2 and Σ) are hyperparameters of this distribution. Point estimation in this setting corresponds to estimating the posterior means of u and β, that is $\mathsf{E}\left[(u, \beta) \mid y; \hat{\sigma}^2, \hat{\Sigma}\right]$.

Conversely, the "likelihood" can be redefined. Once β has been modeled, classically it should be integrated out from the joint distribution to yield the marginal distribution, $p(y; \sigma^2, \Sigma)$. In a sense, the fixed effects become random effects with a known distribution. For the fixed effects there is some tradition in favor of taking this integral - as mentioned the REML estimates can be derived by applying and averaging β over a flat prior. However, little similar precedent exists for the covariance of the random effects, so while it may be ideologically pure to do so, it also may confound expectations.

To be consistent, we instead adopt the perspective of the penalized likelihood, or regularization. For example, putting a prior on β is equivalent to maximizing the posterior density $p(\beta \mid y; \sigma^2, \Sigma)$ to find $p(\hat{\beta} \mid y; \hat{\sigma}^2, \hat{\Sigma})$. While this hybrid quantity may be unusual to the Bayesian, it represents a stop-gap on the way to a full posterior mode obtained by placing priors over all model components. Because **lme4** offers REML estimation, when it makes sense to do so we investigate that integral as well.

A related issue concerning parameterization arises when moving from prior to posterior. Continuing in the perspective of the penalized likelihood, we assume that if a prior is specified on, say, the inverse of the residual variance, the posterior mode of the inverse of the residual variance is desired. In other words, a prior once specified is applied directly as a penalty function, and the Jacobian that results from a change of variables to a canonical form is simply ignored. This can produce some confusion because a parameter may be alternatively reported as a variance and a standard deviation while the posterior mode that is calculated may be of neither. Nevertheless, we find this preferable to finding the posterior modes under implicit transformations and having the awkwardness of a variance not being the square of its associated standard deviation.

3.2. Fixed effect priors

Gaussian priors on the fixed effects require little effort to include as they can be treated as pseudo data. The program for optimization follows almost identical steps as that for the first stage of the likelihood. If we assume that $\beta \mid \sigma^2 \sim \mathcal{N}(0, \sigma^2\Sigma_\beta)$ with Σ_β known, positive definite, and having the decomposition $\Sigma_\beta = L_\beta L_\beta^\top$, then the joint density from equation 3 becomes

$$\begin{aligned} p(y, u, \beta; \sigma^2, \Sigma) &= (2\pi\sigma^2)^{-(N+Q+P)/2}|\Sigma_\beta|^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}\left[\|y - X\beta - Z\Lambda u\|^2 + \|u\|^2 + \beta^\top\Sigma_\beta^{-1}\beta\right]\right\}, \\ &= (2\pi\sigma^2)^{-(N+Q+P)/2}|\Sigma_\beta|^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}\left\|\begin{bmatrix} y \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} Z\Lambda & X \\ 0 & L_\beta^{-1} \\ I_Q & 0 \end{bmatrix}\begin{bmatrix} u \\ \beta \end{bmatrix}\right\|^2\right\}. \end{aligned}$$

Note the occurrence of σ^2 when modeling β. In this case, use of the common scale factor allows us to proceed almost exactly as before: $L_X L_X^\top$ changes from $X^\top X - L_{ZX}L_{ZX}$ to $X^\top X + \Sigma_\beta^{-1} - L_{ZX}L_{ZX}^\top$ and the degrees of freedom for σ^2 increase from N to $N + P$. Once L_X has been redefined, the modes of this function, $\tilde{u}$ and $\tilde{\beta}$, are calculated with no other changes. The marginal posterior used as an objective function for optimization is given by

$$\begin{aligned} p(\beta \mid y; \sigma^2, \Lambda) \propto (\sigma^2)^{-(N+P)/2}|L_Z|^{-1}\times \\ \exp\left\{-\frac{1}{2\sigma^2}\left[(\beta - \tilde{\beta})^\top L_X L_X^\top(\beta - \tilde{\beta}) + \left\|\begin{bmatrix} y \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} Z\Lambda & X \\ 0 & L_\beta^{-1} \\ I_Q & 0 \end{bmatrix}\begin{bmatrix} \tilde{u} \\ \tilde{\beta} \end{bmatrix}\right\|^2\right]\right\}. \end{aligned}$$

The maximum a posteriori estimate of β is $\tilde{\beta}$, and the posterior at its mode has maximal values for the hyperparameter value $\hat{\sigma}^2 = \frac{1}{N+P}\left[\|y - X\hat{\beta} - Z\Lambda\tilde{u}\|^2 + \|\tilde{u}\|^2 + \|\tilde{L}_\beta^{-1}\tilde{\beta}\|^2\right]$. The profiled posterior at its mode is:

$$p(\hat{\beta} \mid y; \hat{\sigma}^2, \Lambda) \propto (\hat{\sigma}^2(\Lambda))^{-(N+P)/2}|L_Z(\Lambda)|^{-1}.$$

Now correctly denoted a likelihood, the REML procedure of integrating out β yields the function:

$$p(y; \sigma^2, \Lambda) = (2\pi\sigma^2)^{-N/2}|\Sigma_\beta|^{-1/2}|L_Z|^{-1}|L_X|^{-1}\exp\left\{-\frac{1}{2\sigma^2}\left\|\begin{bmatrix} y \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} Z\Lambda & X \\ 0 & L_\beta^{-1} \\ I_Q & 0 \end{bmatrix}\begin{bmatrix} \tilde{u} \\ \tilde{\beta} \end{bmatrix}\right\|^2\right\}.$$

The estimator $\hat{\sigma}^2_{\text{RE}}$ is the same as the previous case, except for being scaled by N instead of $N + P$. The profiled likelihood/REML objective function is:

$$p(y; \hat{\sigma}^2, \Lambda) = (2\pi\hat{\sigma}^2_{\text{RE}}(\Lambda))^{-N/2}|\Sigma_\beta|^{-1/2}|L_Z(\Lambda)|^{-1}|L_X(\Lambda)|^{-1}e^{-N/2}.$$

case	function	norm const	ν	S^2
β unmodeled, ML	$p(y;\hat{\beta},\sigma^2,\Lambda)$	$(2\pi)^{-\nu/2}$	N	PRSS
β unmodeled, REML	$q(y;\sigma^2,\Lambda)$	$(2\pi)^{-\nu/2}\lvert L_X\rvert^{-1}$	$N-P$	PRSS
$\beta\sim\mathcal{N}(0,\sigma^2\Sigma_\beta)$, MAP	$p(\hat{\beta}\mid y;\sigma^2,\Lambda)$	NA	$N+P$	PRSS $+\lVert L_\beta^{-1}\beta\rVert^2$
$\beta\sim\mathcal{N}(0,\sigma^2\Sigma_\beta)$, ML	$p(y;\sigma^2,\Lambda)$	$(2\pi)^{-\nu/2}\lvert L_X\rvert^{-1}\lvert\Sigma_\beta\rvert^{-1/2}$	N	PRSS $+\lVert L_\beta^{-1}\beta\rVert^2$

Table 1: The variable components of equation 9 broken down by the choice of model for the fixed effects. 'PRSS' stands for the base-model penalized residual sum of squares, $\|y-X\hat{\beta}-Z\Lambda\tilde{u}\|^2+\|\tilde{u}\|^2$. The first and third rows correspond to straight maximization, while the second and fourth arise from the REML procedure of integrating out β.

Had we instead chosen to incorporate substantive prior knowledge, i.e. assumed $\beta\sim\mathcal{N}(0,\Sigma_\beta)$ without utilizing the common scale, the least-squares problem is solvable but then depends on the value of σ^2. In turn, this needs to be added to the set of parameters optimized numerically. Specifically, the joint density (equation 3) becomes:

$$p(y,u,\beta;\sigma^2,\Sigma)=(2\pi\sigma^2)^{-(N+Q+P)/2}\lvert\Sigma_\beta\rvert^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}\left\|\begin{bmatrix}y\\0\\0\end{bmatrix}-\begin{bmatrix}Z\Lambda & X\\0 & \sigma L_\beta^{-1}\\I_Q & 0\end{bmatrix}\begin{bmatrix}u\\\beta\end{bmatrix}\right\|^2\right\}.$$

$L_XL_X^\top$ would then be computed as factors of $X^\top X+\sigma^2\Sigma_\beta^{-1}-L_{ZX}L_{ZX}^\top$ so that the joint modes $\tilde{u}(\sigma^2,\Lambda)$ and $\tilde{\beta}(\sigma^2,\Lambda)$ are properly written as depending on the common scale as well as the covariance of the random effects.

More complicated priors, such as t distributions, require adding β to the set of parameters for numeric optimization.

3.3. Common scale priors

Under the assumption that the fixed effects can be successfully profiled out without depending on the common scale, we have a first-stage profiled equation of the form

$$f(y;\sigma^2,\Lambda)\propto(\sigma^2)^{-\nu/2}\lvert L_Z\rvert^{-1}\exp\left\{-\frac{1}{2\sigma^2}S^2\right\}. \tag{9}$$

Table 1 enumerates the possible instantiations of this equation for the cases we have considered. Taking a logarithm, derivative with respect to σ^2, and rescaling yields a linear equation with a maximizer in $\hat{\sigma}^2$ that is $\frac{1}{\nu}S^2$.

Unlike the case for the fixed effects, application of an independent prior to σ^2 can proceed directly from this equation without reiterating its derivation. A conjugate prior, that is an inverse gamma on the scale of σ^2, will again yield a linear optimization problem. Specifically, if we use a shape parameter of a and a scale of b, then $\hat{\sigma}^2=\frac{1}{\nu/2+a+1}(S^2/2+b)$.

More complicated results follow from other variants of the gamma family. If we apply a gamma distribution to σ^2, not its inverse, and again with shape a and scale b, we transform the base function objective function into the form

$$g(y; \sigma^2, \Lambda) \propto (\sigma^2)^{-\nu/2+a-1} |L_Z|^{-1} \exp\left\{ -\frac{1}{2\sigma^2} S^2 - \frac{\sigma^2}{b} \right\}.$$

A logarithm and a derivative with respect to σ^2 produces a quadratic problem. Taking the non-negative square root yields

$$\hat{\sigma}^2 = \frac{b}{2} \left[\sqrt{(\nu/2 - a + 1)^2 + 2S^2/b} - (\nu/2 - a + 1) \right].$$

It is also possible to apply an inverse gamma to σ as a standard deviation and obtain a quadratic problem. The objective function and its mode under a gamma prior with shape a and scale b are given by:

$$g(y, \sigma^2, \Lambda) = (\sigma^2)^{-(\nu+a+1)/2} \exp\left\{ -\frac{1}{2\sigma^2} S^2 - \frac{b}{\sigma} \right\},$$
$$\hat{\sigma} = \frac{b + \sqrt{b^2 + 4(\nu + a + 1)S^2}}{2(\nu + a + 1)}.$$

Finally, we also note that it is possible, and often desirable, to use a point-mass prior for the residual variance. For example, this arises in meta-analyses which can be written as weighted linear mixed models having a residual variance of 1. While maintaining the same general profiling procedure, little effort is required to simply plug-in a specific value for σ^2 instead of estimating it.

In summary, for the priors on β that have maximizers independent of σ^2, the application an a) inverse-gamma prior on σ^2, b) gamma on σ^2, or c) inverse-gamma on σ yield linear, quadratic, or quadratic optimization problems respectively. Point priors require no optimization at all. When these conditions are satisfied we proceed as before, first solving for the maximum in (u, β) as before, then using these to find the maximum in σ^2 if necessary, and finally by plugging these values into a likelihood/objective function such as equation 5. The maximizer in σ^2 may no longer be proportional to the exponential term so that the resulting equation may have a more complex form than, say the profiled likelihood (equation 6), however this does not come with increased computational complexity.

Finally we note that it is also possible to apply more complicated priors and maintain the profiling scheme so long as a unique mode for σ^2 exists. An inner-loop single parameter optimization or, equivalently, a root finding algorithm can serve the role of explicit derivations without greatly increasing running time. The final case considered in section 3.2, with a prior for the fixed effects having a covariance specified on an absolute scale, can also be computed using single parameter optimization at this step. However, doing so requires that the least-squares problem be solved for every change in value of σ^2.

3.4. Random effect covariance priors

The first concern that arises when placing a prior over the covariance of the random effects is the structure of that matrix. While we have consistently written simply 'Σ' as if it were a monolithic construct, in reality this matrix consists of block repetitions of individual submatrices along its diagonal. If there are K grouping factors with q_k varying coefficients

and J_k different groups at level k, then we have $\Sigma_1, \ldots, \Sigma_k$ distinct covariance matrices and $\Sigma = \text{diag}_{k=1}^{K}(I_{J_k} \otimes \Sigma_k)$. In words, Σ consists of Σ_1 repeated J_1 times along the diagonal, then Σ_2 repeated J_2 times, and so forth. A prior over Σ is in reality a prior over $\Sigma_1, \ldots, \Sigma_k$, or equivalently a reparameterization of these matrices. We denote the free parameters of these matrices as θ. For simplicity, we only consider covariance priors that factor independently across these submatrices, although joint priors remain an avenue for future work.

An examination of the profiled likelihood (equation 6) or REML objective function (equation 8) demonstrates that priors on Σ independent of the other parameters can be applied directly to the covariance of the random effects - divided by the residual variance - by simply adding a penalty term. For example, in the profiled posterior

$$\begin{aligned} p(\Sigma(\theta) \mid y; \hat{\beta}, \hat{\sigma}^2) &\propto p(y; \hat{\beta}, \hat{\sigma}^2, \Sigma(\theta)) p(\Sigma(\theta)), \\ &\propto \left\{ \hat{\sigma}^2(\theta)^{-N/2} |L_Z(\theta)|^{-1} \right\} p(\Sigma(\theta)), \end{aligned}$$

the first term on the right-hand side, the profiled likelihood, is calculated exactly as before.

The situation is more complicated when the prior is to be applied in an absolute sense, that is not scaled by the residual variance. In particular, this situation arises when substantive prior information exists. As will be shown, imposing a prior of this sort that is independent of the other parameters is actually equivalent to placing a joint prior on σ^2 and Σ. If profiling is to be preserved, covariance priors not on the common scale are restricted to cases similar to those outlined for the residual variance. It is important to reiterate that the discussion that follows is entirely limited to this case, and that scale-free optimization is trivial in comparison.

As an example of the problems involved, consider a set of data with a single grouping factor. For notational simplicity, we will drop the $k = 1$ subscript when possible so that $q = q_1$ is the number of varying coefficients at this level and $\tilde{\Sigma} = \sigma^2 \Sigma_1$ is the associated absolute-scale covariance matrix. Suppose that it is desired to place an inverse Wishart prior on $\tilde{\Sigma}$ with μ degrees of freedom and scale matrix Ψ. Applying this to the likelihood (equation 5) produces the posterior density

$$\begin{aligned} p(\tilde{\Sigma} \mid y; \beta, \sigma^2) \propto (\sigma^2)^{-N/2} |L_Z|^{-1} |\tilde{\Sigma}|^{-(\mu+q+1)/2} \exp\left\{ -\frac{1}{2 \cdot 1} \text{tr}\left(\Phi \tilde{\Sigma}^{-1} \right) \right\} \times \\ \exp\left\{ -\frac{1}{2\sigma^2} \left[(\beta - \tilde{\beta})^\top L_X L_X^\top (\beta - \tilde{\beta}) + \left\| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} Z\Lambda & X \\ I_Q & 0 \end{bmatrix} \begin{bmatrix} u \\ \beta \end{bmatrix} \right\|^2 \right] \right\}. \end{aligned}$$

However, this equation is to be optimized point-wise. For the values of $\hat{\beta}$, $\hat{\sigma}^2$, and $\hat{\tilde{\Sigma}}$ that maximize it, $\hat{\beta}$, $\hat{\sigma}^2$, and $\hat{\Sigma}_1 = \hat{\tilde{\Sigma}}/\hat{\sigma}^2$ achieve the same value. That is, the right-hand side is the same if we instead work with the equation:

$$\begin{aligned} p(\tilde{\Sigma} \mid y; \beta, \sigma^2) \propto (\sigma^2)^{-(N+\mu+q+1)/2} |L_Z|^{-1} |\Sigma_1|^{-(\mu+q+1)/2} \times \\ \exp\left\{ -\frac{1}{2\sigma^2} \left[(\beta - \tilde{\beta})^\top L_X L_X^\top (\beta - \tilde{\beta}) + \left\| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} Z\Lambda & X \\ I_Q & 0 \end{bmatrix} \begin{bmatrix} u \\ \beta \end{bmatrix} \right\|^2 + \text{tr}\left(\Phi \Sigma_1^{-1} \right) \right] \right\}. \end{aligned}$$

Consequently, we have that an inverse Wishart prior on the unscaled matrix $\tilde{\Sigma}$ with degrees of freedom μ and scale Ψ is equivalent to the joint prior

$$\begin{aligned}\sigma^2 \mid \Sigma_1 &\sim \text{Inv} - \text{Gamma}\left(\text{shape} = (\mu + q - 1)/2,\ \text{scale} = \text{tr}\left(\Psi\Sigma_1^{-1}\right)/2\right), \\ \Sigma_1 &\sim \text{Inv} - \text{Wishart}(\mu, \Psi).\end{aligned}$$

Doing similar calculations for the cases of a Wishart prior on $\tilde{\Sigma}$ and an inverse Wishart on the unique positive definite square root of $\tilde{\Sigma}$ yield analogous schemes to the optimization of σ^2 in section 3.3, with an induced gamma distribution on σ^2 and inverse gamma distribution on σ respectively.

Taking into consideration these induced priors and the added complexity of Σ being comprised of multiple covariance sub-matrices, straightforward profiling of σ^2 is constrained to the cases where the totality of all induced priors yield a quadratic or linear optimization problem. Specifically, inverse Wishart distributions applied to the unscaled covariance matrices can be added in any quantity desired, as the net effect is to add terms to the penalized residual sum of squares. For any remaining grouping factors for which an absolute-scaled prior is desired, either Wishart distributions on covariance matrixes or inverse Wishart distributions on covariance square roots must be chosen and applied consistently. To profile σ^2, the various coefficients of the exponential terms are collected for each optimization iteration by cycling through the grouping factors, and the resulting linear or quadratic problem solved.

Finally, we briefly discuss improper Wishart-family priors. The principle complication in modeling unscaled random effect covariances is in the exponential term involving σ^2 introduced into the likelihood. The polynomial term simply increases or decreases the degrees of freedom. Thus, improper distributions that eliminate the exponential term are trivial to include. These are the limits of gamma/Wishart distributions as the scale tends to infinity or the limits of the inverse as the scale tends to 0. Further allowing improper degrees of freedom as well shows that the choice of prior parameterization is irrelevant, that is Wishart priors on covariances, square roots, or their inverses all yield penalties terms that are powers of the determinant.

3.5. Summary

Table 2 lists the priors that in our discussion we have shown can be fully profiled down to functions that are just the covariance of the random effects. Provided that caution is exercised so that the optimization of the residual variance is straightforward, an option can be selected from each column. In all cases, the algorithm to calculate the profiled objective function is given by:

1. Determine the maximizer of the joint density of the observations and the spherical random effects in (u, β). In doing so, the integral of the joint density with respect to the random effects is effectively computed as well.

2. Plug in the joint mode of β to the likelihood, as it is also the maximum likelihood/maximum a posteriori estimate.

Fixed Effects β	Residual Variance σ^2	Random Effect Covariance Σ_k
$\beta \mid \sigma^2 \sim \mathcal{N}(0, \sigma^2\Sigma_\beta)$	$\sigma^2 \sim \text{Inv} - \text{Gamma}$	$p(\Sigma_k)$ arbitrary
	$\sigma^2 \sim \text{Gamma}$	
	$\sigma \sim \text{Inv} - \text{Gamma}$	$\tilde{\Sigma}_k = \sigma^2\Sigma_k$,
	$\sigma^2 = \sigma_0^2$ w.p.1	$\tilde{\Sigma}_k \overset{\text{ind}}{\sim} \text{Wish/Inv} - \text{Wish}$
		$\tilde{\Sigma}_k^{1/2} \overset{\text{ind}}{\sim} \text{Inv} - \text{Wish}$

Table 2: Priors for linear mixed models that can be fit into the two-stage profiling scheme of **lme4**.

3. Calculate the mode of σ^2 as the solution to a linear or quadratic problem, depending on the choice of priors. Point priors involve no optimization, and more complicated cases may require root-finding.

4. Plug in the estimate of σ^2 and numerically optimize the resulting function.

4. Generalized linear mixed model specification

As there are no obvious ways by which a generalized linear mixed model can be profiled, **lme4** performs optimization for these models numerically across the entire set of parameters. In order to apply priors on these models as well, we write out their specification. Generalized linear models principally differ from their regular counterparts through the use of a link function, here denoted 'g', that maps linear combinations of coefficients to the expected value of the response. In addition, the response given the random effects is no longer assumed Gaussian but instead merely a member of the the exponential family of distributions. While omitting the specifics of this conditional distribution, generalized linear mixed models can be written as

$$\begin{aligned} \mathsf{E}[y \mid b; \beta, \Sigma] &= g^{-1}(X\beta + Zb), \\ b &\sim \mathcal{N}(0, \Sigma), \end{aligned}$$

where g^{-1} is the inverse of the link function and it is to be understood as applied component-wise to the vector of linear predictors.

Given this specification, the parameters of the model are the fixed effects, β, and the covariance of the random effects, Σ. Like linear mixed models, Σ is highly structured and comprised of sub-matrices. Unlike linear mixed models, there is no scale parameter/residual variance. The family associated with the response may have additional parameters, but as of this writing the "quasi" options that include overdispersion are not supported by **lme4**.

lme4 performs optimization for generalized linear mixed models by assuming values for β and Σ, approximating the integral over the random effects, and returning a value for the likelihood to a numeric optimizer. To apply a prior to either parameter, it is sufficient to include a penalty term at this last step.

5. blme overview

We have written the software package **blme** for the R statistical programming language that performs profiled posterior maximization in linear and generalized linear mixed models. **blme** is based on **lme4** and uses as much machinery of that package possible. In particular, the efficient C++ code and matrix decompositions that underly an **lme4** model fit are also used by **blme**.

5.1. Calling blme functions

blme was designed to be familiar to users of **lme4**. A `bmerMod` S4 object extends the `merMod` class, and consequently inherits all of the same functionality. In **lme4**, linear mixed models are fit using the `lmer` function, while generalized models use `glmer`. Fitting a model in **blme** is achived by instead calling `blmer` or `bglmer`, modified versions of the original functions with new arguments.

The prototypes for `blmer` and `bglmer` are:

```
blmer(formula, data, REML = TRUE, control = lmerControl(), start = NULL,
      verbose = 0L, subset, weights, na.action, offset, contrasts = NULL,
      devFunOnly = FALSE,
      cov.prior = wishart, fixef.prior = NULL, resid.prior = NULL, ...)

bglmer(formula, data, family = gaussian, control = glmerControl(),
       start = NULL,  verbose = 0L, nAGQ = 1L, subset, weights, na.action,
       offset, contrasts = NULL, mustart, etastart, devFunOnly = FALSE,
       cov.prior = wishart, fixef.prior = NULL, ...)
```

For both, all but the last line are identical to their **lme4** equivalent. The new arguments are `cov.prior`, `fixef.prior`, and `resid.prior`. All three apply to linear mixed models, while only the first two apply to generalized ones, for the reasons discussed in section 4.

The format for each new argument is analogous to that of a delayed function call that is evaluated in a special environment. This is done so that priors can refer to particulars of the model without having these explicitly defined at the calling level, for example allowing a default prior for random effect covariance matrices that scales with the dimensions of the grouping factors. Variables available at prior specification are:

1. `q.k` or `level.dim` - dimension of Σ_k, or how many coefficients vary at the kth level; covariance priors only
2. `j.k` or `n.grps` - number of groups at the kth level; covariance priors only
3. `n` or `n.obs` - number of observations
4. `p` or `n.fixef` - number of fixed effects

Exceptions to the function-call syntax are that 0-argument function invocations can be expressed without the empty `()` parentheses, character strings can be passed instead to aid

in reuse, and for multiple grouping factors R formula notation is abused to specify different priors for different levels. In the case of multiple priors for a single argument, the input should be a list. Examples and descriptions follow in the relevant sections below.

Finally, note that the default random effect covariance prior is set to `wishart` for both functions. This provides a weakly-informative prior that mitigates the downward bias of covariance estimates and guarantees that the result will be strictly positive definite TODO: REF. For the sake of exposition, in the fixed effect and conditional prior examples discussed below the covariance prior will be explicitly disabled so that model being discussed differs from the **lme4** version by only the prior under consideration.

5.2. Fixed effect priors

Currently supported fixed effect priors are flat/`NULL`, multivariate normal, and multivariate Student's t distributions. Flat and normal priors can be profiled without adding the fixed effects to the numeric optimization parameter set. t distributions require additional computation, and thus can be considerably slower to fit. In addition, t priors cannot be used when the `REML` is true. See section 3.2 for details.

To specify a fixed effect prior, pass to `blmer` or `bglmer` a value for the `fixef.prior` argument of the format:

```
distribution.name(options.list)
```

where `distribution.name` can be `normal` or `t`. An argument of `NULL`/`flat` is equivalent to a flat prior or no penalty term. Options and defaults are:

1. `normal(sd = c(10, 2.5), cov, common.scale = TRUE)`
   (a) `sd` - a vector of standard deviations. If length 1, the value is reused for all components. If length 2, the first item applies to the first fixed effect (typically the intercept term), while the second is reused. Otherwise must be of length equal to the number of fixed effects.
   (b) `cov` - either a vector of variances of length equal to the number of fixed effects or a positive definite matrix of appropriate size. Only one of `sd` or `cov` should be specified.
   (c) `common.scale` - a logical that when true implies that prior is of the form $\beta \mid \sigma^2 \sim \mathcal{N}(0, \sigma^2\Sigma_\beta)$. When false, $\beta \sim \mathcal{N}(0, \Sigma_\beta)$. Only applies to linear mixed models.
2. `t(df = 3, scale = c(10^2, 2.5^2), common.scale = TRUE)`
   (a) `df` - a positive scalar signifying the degrees of freedom.
   (b) `scale` - determines the scale matrix. If is a scalar, the scale is that value times the identity matrix. For a length of 2 the second value is placed along the diagonal for all intercepts while the off-diagonals are set to 0. For a length equal to the number of fixed effects, the values are placed directly on the diagonal. Otherwise, a positive definite matrix of the appropriate size can be given.
   (c) `common.scale` a logical that when true implies that prior is of the form $\beta \mid \sigma^2 \sim t_\nu(0, \sigma^2\Sigma_\beta)$. When false, $\beta \sim t_\nu(0, \Sigma_\beta)$. Only applies to linear mixed models.

The t prior used has the density

$$\beta \sim t_\nu(0, \Sigma_\beta) \Rightarrow p(\beta) = \frac{\Gamma\left((\nu + P)/2\right)}{\Gamma(\nu/2)(\nu\pi)^{P/2}|\Sigma_\beta|^{1/2}\left[1 + \frac{1}{\nu}\beta^\top \Sigma_\beta^{-1}\beta\right]^{(\nu+P)/2}},$$

where P is the number of fixed effects. TODO: reference.

5.3. Fixed effect examples

For the purposes of illustration, suppose that we have defined in the global environment a response variable `y`, a predictor `x`, and a grouping factor `g`.

Default normal prior

The following applies a normal prior using the default hyperparameters, while disabling the default random effect covariance prior.

```
blmer(y ~ 1 + x + (1 + x | g), cov.prior = NULL,
      fixef.prior = normal)
```

Normal prior with specific scales

To do ridge regression, one can cross-validate to find the optimal `lambda`:

```
blmer(y ~ 1 + x + (1 + x | g), cov.prior = NULL,
      fixef.prior = normal(sd = 1 / sqrt(lambda)))
```

Normal prior with substantive information

With prior information for the scale of the fixed effects, it becomes necessary to model β without using the common scale. Assuming that this information is in the matrix `Sigma.beta`, the `blmer` call to incorporate it is:

```
blmer(y ~ 1 + x + (1 + x | g), cov.prior = NULL,
      fixef.prior = normal(cov = Sigma.beta, common.scale = FALSE))
```

t prior for logistic regression

Gelman *et al.* (2008) recommend a Cauchy prior for logistic regression. `bglmer` fits an analogous mixed model by using an appropriately parameterized t distribution:

```
bglmer(y ~ 1 + x + (1 + x | g), cov.prior = NULL,
       family = binomial,
       fixef.prior = t(df = 1, scale = 2.5^2))
```

5.4. Residual variance priors

A prior on the residual variance/common scale is specified in the same format as for fixed effects, but instead using the `resid.prior` argument to `blmer`. As discussed in section 4, the generalized linear mixed models fit by **lme4** do not have a scale parameter.

The named distributions that can be used are: `gamma`, `invgamma`, `point`, and `NULL`/`flat`. Options and defaults are:

1. `gamma(shape = 0, rate = 0, posterior.scale = "var")`
   (a) `shape` - non-negative scalar. For a shape of 0, the prior is improper.
   (b) `rate` - non-negative scalar. For a rate of 0, the prior is improper.
   (c) `posterior.scale` - one of `"sd"` or `"var"`, corresponding to $\sigma \sim \Gamma(\text{shape}, \text{rate})$ and $\sigma^2 \sim \Gamma(\text{shape}, \text{rate})$ respectively.

   The default setting applies the improper prior $p(\sigma^2) \propto (\sigma^2)^{-1}$.
2. `invgamma(shape = 0, scale = 0, posterior.scale = "var")`
   (a) `shape` - non-negative scalar. For a shape of 0, the prior is improper.
   (b) `scale` - non-negative scalar. For a scale of 0, the prior is improper.
   (c) `posterior.scale` - one of `"sd"` or `"var"`, having the same interpretation as for `gamma`.
3. `point(value = 1, posterior.scale = "sd")`
   Fixes σ^2 to a specific value, and is equivalent to a point-mass prior.
   (a) `value` - a positive scalar.
   (b) `posterior.scale` - one of `"sd"` or `"var"`, having the same interpretation as for `gamma`.

5.5. Residual variance example

A common use for residual variance priors is when the observations have known but unequal residual variances. Assuming that these are defined in the variable `resid.var`, this model can be fit in `blmer` by the call:

```
blmer(y ~ 1 + x + (1 + x | g), cov.prior = NULL,
      weights = 1 / resid.var,
      resid.prior = point)
```

5.6. Random effect covariance priors

As with a mixed effect model there can be multiple grouping factors each with its own distinct covariance matrix, **blme** supports the specification of a default prior as well as priors specific to named levels. To apply a prior to single level, the input should be of the form:

```
grouping.name ~ distribution.name(options.list)
```

Conversely, to specify a default prior that applies to all grouping factors, the format is simply: `distribution.name(options.list)`. The distributions and their options are:

1. `gamma(shape = 2.5, rate = 0, common.scale = TRUE, posterior.scale = "sd")` Can only be used with univariate grouping factors.
   (a) `shape` - a non-negative scalar. For a shape of 0, the prior is improper.
   (b) `rate` - a non-negative scalar. For a rate of 0, the prior is improper.
   (c) `common.scale` - a logical that when true indicates that the prior should be applied to the random effect covariance divided by the residual variance while when false the prior is applied to the covariance in the absolute sense. Note that this is slightly different than the interpretation for fixed effects, as there it influenced the scale and not the parameters themselves. Only applies to linear mixed models. See sections 2.2 and 3.4.
   (d) `posterior.scale` - one of `"sd"` or `"var"` corresponding to $\sqrt{\Sigma_k} \sim \Gamma(\text{shape}, \text{rate}$ and $\Sigma_k \sim \Gamma(\text{shape}, \text{rate}$ respectively.
2. `invgamma(shape = 0.001, scale = shape + 0.05, common.scale = TRUE, posterior.scale = "var")`
   (a) `shape` - a non-negative scalar. For a shape of 0, the prior is improper.
   (b) `scale` - a non-negative scalar. For a scale of 0, the prior is improper.
   (c) `common.scale` - logical with similar interpretation as in the `gamma` case.
   (d) `posterior.scale` - one of `"sd"` or `"var"` having a similar interpretation as for `gamma`.
3. `wishart(df = level.dim + 2.5, scale = Inf, common.scale = TRUE, posterior.scale = "cov")`
   (a) `df` - a scalar greater than or equal to the number of varying coefficients for a level, minus 1. When equal to the lower bound, the prior is improper.
   (b) `scale` - a single value, a vector of length equal to the grouping factor dimension, or an appropriately sized positive definite matrix. For the first case, the value is multiplied by the identity matrix; for the second, the vector is made into a diagonal matrix. `Inf` is allowed, and yields an improper distribution.
   (c) `common.scale` - logical with similar interpretation as in the `gamma` case.
   (d) `posterior.scale` - one of `"sqrt"` or `"var"` having a similar interpretation as for `gamma`. Note that the unique matrix square root can be expensive to compute, as the optimization is not performed in this parameterization.
4. `invwishart(df = level.dim - 0.998, scale = diag(df + 0.1, level.dim), common.scale = TRUE, posterior.scale = "cov")`
   (a) `df` - a scalar greater than or equal to the number of varying coefficients for a level, minus 1. When equal to the lower bound, the prior is improper.

(b) `scale` - a single value, a vector of length equal to the grouping factor dimension, or an appropriately sized positive definite matrix. For the first case, the value is multiplied by the identity matrix; for the second, the vector is made into a diagonal matrix. `0` is allowed, and yields an improper distribution.

(c) `common.scale` - logical with similar interpretation as in the `gamma` case.

(d) `posterior.scale` - one of `"sqrt"` or `"var"` having a similar interpretation as for `wishart`.

5. `custom(fn, chol = FALSE, common.scale = TRUE, scale = "none")`
Can be used to apply an arbitrary function as a penalty

(a) `fn` - function to be used as a prior. Must take an argument as specified by `chol` and return a value as specified by `scale`.

(b) `chol` - a logical indicating whether or not `fn` expects a covariance matrix (`FALSE`) or a *right* Cholesky factor (`TRUE`).

(c) `common.scale` - logical with same interpretation as other cases. When `FALSE`, it is unknown if the residual variance can be profiled and is thus is added to the set of parameters for numeric optimization.

(d) `scale` - one of `"none"`, `"log"`, or `"dev"` corresponding to `fn` returning values on the scales $p(\Sigma)$, $\log p(\Sigma)$, and $-2 \log p(\Sigma)$ respectively.

5.7. Covariance examples

For the following, we now assume that there are two grouping factors, `g.1` and `g.2`.

lme4 fit

By default, **blme** applies a Wishart prior to the covariance of the random effects. By suppressing this, the following two calls are equivalent:

```
lmer (y ~ 1 + x + (1 | g.1) + (1 + x | g.2))
blmer(y ~ 1 + x + (1 | g.1) + (1 + x | g.2),
      cov.prior = NULL)
```

Univariate, default prior, standard deviation scale

The follow places a prior on the standard deviation of the contributions to the intercept for the first grouping factor:

```
blmer(y ~ 1 + x + (1 | g.1) + (1 + x | g.2),
      cov.prior = gamma)
```

As the prior was univariate, it does not extend to the second grouping factor. Instead, this remains unmodeled. If a third level with a single varying coefficient existed, the gamma prior would apply to that as well.

Multivariate, default prior, variance scale

Using a Wishart as a default specializes down to a gamma for a univariate case and consequently applies to both grouping factors. The default for the Wishart is to use a posterior scale of a covariance, unlike the previous example.

```
blmer(y ~ 1 + x + (1 | g.1) + (1 + x | g.2),
      cov.prior = wishart)
```

Named grouping factors

We can mix the above examples by naming the grouping factors.

```
blmer(y ~ 1 + x + (1 | g.1) + (1 + x | g.2),
      cov.prior = list(g.1 ~ gamma, g.2 ~ wishart))
```

Default priors with options

As with univariate families there are no complications with level dimensions, it is easy to specify a default that applies in more than one case. For the following, we no longer model the random effects of the second grouping factor as having a varying slope.

```
blmer(y ~ 1 + x + (1 | g.1) + (1 | g.2),
      cov.prior = gamma(shape = 3, rate = 1, posterior.scale = 'var'))
```

Fixed hyperparameters

The parent of the evaluating environment for the prior creation functions is set to the one which calls `blmer` so it is possible to use variables defined there.

```
g.2.cov <- matrix(c(1, 0.2, 0.2, 1), 2)
blmer(y ~ 1 + x + (1 | g.1) + (1 + x | g.2),
      cov.prior = g.2 ~ wishart(scale = g.2.cov))
```

Complex expressions

When used with default settings, the Wishart has its degrees of freedom set to the number of coefficients varying at a grouping factor plus 2.5. The following applies a default prior that penalizes the covariance for each level by a term that is $|\Sigma_k|^{1/2}$. This results in a linear term on the standard deviation when there is only one varying coefficient.

```
blmer(y ~ 1 + x + (1 | g.1) + (1 + x | g.2),
      cov.prior = wishart(df = level.dim + 2))
```

For a Wishart distribution with ν degrees of freedom, q_k dimension, and a scale matrix of Ψ, its mode is given by $(\nu - q_k - 1)\Psi$. We can thus construct a proper verison of the Wishart prior that is parameterized by its degrees of freedom and mode and use this as a default.

```
prior.mode <- 1e-4
blmer(y ~ 1 + x + (1 | g.1) + (1 + x | g.2),
      cov.prior = wishart(scale = diag(prior.mode, q.k) / (df - q.k - 1)))
```

Custom prior

TODO: REF recommends a half-Cauchy prior for variance components in linear mixed models. For univariate grouping factors,

```
penaltyFn <- function(sigma) dcauchy(sigma, 0, 10, log = TRUE)
blmer(y ~ 1 + x + (1 | g.1) + (1 | g.2),
      cov.prior = custom(penaltyFn, chol = TRUE, scale = "log"))
```

5.8. `blme` output

blme functions return objects that inherit from the **lme4** base classes. For linear models, the result is a `blmerMod` derived from `lmerMod`, while in the general case the result is of class `bglmerMod`, derived from `glmerMod`.

The `summary` function for each highlights the main differences from **lme4**.

```
> summary(fm1 <- lmer(Reaction ~ Days + (Days | Subject), sleepstudy))

Linear mixed model fit by REML ['lmerMod']
Formula: Reaction ~ Days + (Days | Subject)
   Data: sleepstudy

REML criterion at convergence: 1743.6

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.9536 -0.4634  0.0231  0.4634  5.1793

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 612.09   24.740
          Days         35.07    5.922   0.07
 Residual             654.94   25.592
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
```

```
> summary(bm1 <- blmer(Reaction ~ Days + (Days | Subject), sleepstudy))

Cov prior  : Subject ~ wishart(df = 4.5, scale = Inf, posterior.scale = cov,
           : common.scale = TRUE)
Prior dev  : 2.4021

Linear mixed model fit by REML ['blmerMod']
Formula: Reaction ~ Days + (Days | Subject)
   Data: sleepstudy

REML criterion at convergence: 1749.5

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.2393 -0.4485 -0.0003  0.4858  5.3884

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 840.64   28.994
          Days         97.29    9.864   -0.30
 Residual             606.69   24.631
Number of obs: 180, groups:  Subject, 18

Fixed effects:
            Estimate Std. Error t value
(Intercept)  251.405      7.638   32.91
Days          10.467      2.411    4.34

Correlation of Fixed Effects:
     (Intr)
Days -0.362
```

Aside from the different point estimates, the first lines of output a) provide a textual summary of the priors used and b) quantify the contribution of the priors to the objective function. This information can also be accessed directly from the fitted result.

```
> bm1@priors

$covPriors
$covPriors[[1]]
An object of class "bmerWishartDist"
Slot "df":
[1] 4.5

Slot "R.scale.inv":
     [,1] [,2]
[1,]    0    0
```

```
[2,]    0    0

Slot "log.det.scale":
[1] Inf

Slot "posteriorScale":
[1] "cov"

Slot "commonScale":
[1] TRUE
```

```
> round(bm1@devcomp$cmp, 3)
```

```
      ldL2      ldRX2       wrss       ussq      pwrss      drsum       REML
    96.586      7.130  95968.324  12022.569 107990.893         NA   1749.486
       dev    sigmaML  sigmaREML    penalty
        NA     24.494     24.631      2.402
```

This enables a direct comparison of the distance of the fit from the maximum likelihood or REML estimate.

6. Examples

6.1. Complete separation for mixed models with binary outcomes

McKeon *et al.* (2012) report an experiment on the frequency of predation of coral reefs and the effect of cohabiting crustaceans who defending the colony. The data consist of four treatment groups corresponding to combinations of symbiotic species, and for each treatment pairs of observations are made across 10 temporal blocks. Exploring the data, we find:

```
> culcita[1:5,]
```

```
  block predation  ttt
1     1         0 none
2     1         1 none
3     2         1 none
4     2         1 none
5     3         1 none
```

```
> table(culcita$block)
```

```
 1  2  3  4  5  6  7  8  9 10
 8  8  8  8  8  8  8  8  8  8
```

```
> levels(culcita$ttt)
```

```
[1] "none"   "crabs"  "shrimp" "both"
```

Using `glmer` we can fit a generalized linear mixed model with binomial response and a logistic link function that allows for a random effect corresponding to the temporal block. Following Gelman *et al.* (2008), we standardize binary inputs by centering them to have a mean of 0 and store the result in `culcita.z`. For compactness, we use the `display` function from the **arm** package to show the fit.

```
> m1 <- glmer(predation ~ tttcrabs + tttshrimp + tttboth + (1 | block),
+             culcita.z, family = binomial)
> display(m1)

glmer(formula = predation ~ tttcrabs + tttshrimp + tttboth +
    (1 | block), data = culcita.z, family = binomial, nAGQ = 10)
            coef.est coef.se
(Intercept)  1.60     1.21
tttcrabs    -3.75     1.38
tttshrimp   -4.36     1.45
tttboth     -5.55     1.52

Error terms:
 Groups   Name        Std.Dev.
 block    (Intercept) 3.50
 Residual             1.00
---
number of obs: 80, groups: block, 10
AIC = 70.3, DIC = 60.3
deviance = 60.3
```

A closer look at the data reveals that the model's stability hinges on the presence of a single observation. In fact, in the control group there are only two observations that fail to exhibit predation, one in the first block and one in the 10th.

```
> with(culcita, sum(predation == 0 & ttt == "none"))

[1] 2
```

If for the control no observations were positive, then the likelihood could be made arbitrarily large by increasing the value of the control group's coefficient, *i.e.* quasi-complete separation. On the other hand, in the first temporal block only one observation shows predation. Furthermore, it contains one of the two observations that are preventing separation.

```
> subset(culcita, block == 1)

   block predation    ttt
1      1         0   none
```

```
2      1         1   none
21     1         0 shrimp
22     1         0 shrimp
41     1         0  crabs
42     1         0  crabs
61     1         0   both
62     1         0   both
```

These two circumstances combine so that removing the negative case for the control group in the 10th temporal block nearly makes the problem ill-posed. For consistency, we also remove its pair and re-standardize the result in the variable `culcitaSep.z`. The fitted generalized linear mixed model yields substantially different estimates.

```
> m2 <- glmer(predation ~ tttcrabs + tttshrimp + tttboth + (1 | block),
+             culcitaSep.z, family = binomial)
> display(m2)

glmer(formula = predation ~ tttcrabs + tttshrimp + tttboth +
    (1 | block), data = culcitaSep.z, family = binomial, nAGQ = 10)
            coef.est coef.se
(Intercept)   5.19     4.22
tttcrabs    -15.47    12.63
tttshrimp   -17.46    12.60
tttboth     -20.50    12.56

Error terms:
 Groups   Name        Std.Dev.
 block    (Intercept) 11.75
 Residual              1.00
---
number of obs: 78, groups: block, 10
AIC = 49.1, DIC = 39.1
deviance = 39.1
```

For example, the fixed effect for the intercept jumped from 1.6 to 5.19 corresponding to a change in baseline probability of 0.83 to 0.99, most of the reported statistical signifiance is lost, and the standard deviation of the random effects went from 3.5 to 11.75.

While this is not strictly quasi-complete separation, the situation is similar. The joint distribution of the observations and the random effects can be made large by making the intercept a large positive value and the random effect for the first block highly negative. The extent to which this is possible depends on the covariance of the random effects, which in turn is inflated. The model only remains identifiable because the covariance cannot increase indefinitely.

While not as good as having the full data, with `bglmer` it is possible to at least obtain sensible estimates. Gelman *et al.* (2008) recommends the use of a Cauchy prior with a scale between 0.75 and 2.5 for non-hierarchical logistic regressions. With the knowledge that the problem is biased towards large parameter values, we opt for the smaller scale parameter.

```
> m3 <- bglmer(predation ~ tttcrabs + tttshrimp + tttboth + (1 | block),
+              culcitaSep.z, family = binomial,
+              cov.prior = NULL, fixef.prior = t(1, 0.75))
> display(m3)

bglmer(formula = predation ~ tttcrabs + tttshrimp + tttboth +
    (1 | block), data = culcitaSep.z, family = binomial, nAGQ = 10,
    cov.prior = NULL, fixef.prior = t(1, 0.75))
            coef.est coef.se
(Intercept)  1.88     1.51
tttcrabs    -4.07     1.53
tttshrimp   -4.81     1.64
tttboth     -6.32     1.78

Error terms:
 Groups   Name        Std.Dev.
 block    (Intercept) 4.40
 Residual             1.00
---
number of obs: 78, groups: block, 10
AIC = 53.9, DIC = 43.9
deviance = 43.9
```

The point estimates are all still greater in magnitude than the maximum likelihood ones for the full data, but are now on the same scale. This can be visualized in figure 2, which shows the profiled objectives for the three models as functions of the standard deviation of the random effects alone.

7. References

References

Gelman A, Jakulin A, Pittau MG, Su YS (2008). "A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models." *The Annals of Applied Statistics*, **2**(4), 1360–1383. doi:10.1214/08-AOAS191. URL http://www.jstor.org/stable/30245139.

McKeon CS, Stier AC, McIlroy SE, Bolker BM (2012). "Multiple defender effects: synergistic coral defense by mutualist crustaceans." *Oecologia*, **169**(4), 1095–1103. ISSN 0029-8549. doi:10.1007/s00442-012-2275-2. URL http://dx.doi.org/10.1007/s00442-012-2275-2.

8. Appendix

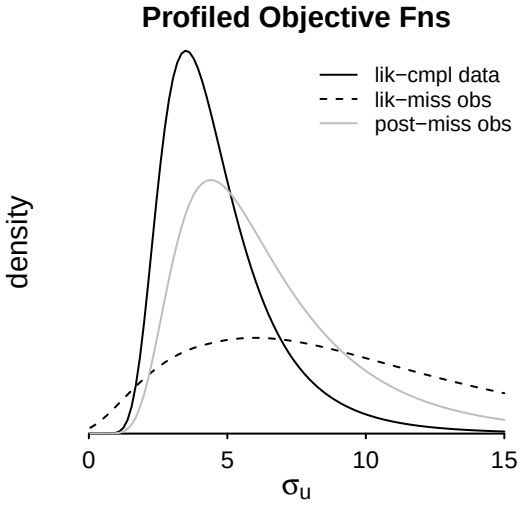

Figure 2: Objective functions for the three models of 6.1. The x axis corresponds to the standard deviation of the random effects, denoted σ_u. Fixed effects have been profiled out numerically. For the likelihood with complete data (`m1`) and the posterior with a Cauchy prior on the fixed effects but incomplete data (`m3`), the curves rescaled so that they roughly integrate to 1. For `m2`, the rescaling was chosen to facilitate visual comparison.

8.1. The joint mode in u and β

Here we write out expressions to find the mode of the joint density of y and u (equation 3) in u and β. If we write $z = \begin{bmatrix} y \\ 0 \end{bmatrix}$, $V = \begin{bmatrix} Z\Lambda & X \\ I_Q & 0 \end{bmatrix}$, and $\gamma = \begin{bmatrix} u \\ \beta \end{bmatrix}$ then we would have a normal equations problem with the maximizer $\hat{\gamma} = (V^\top V)^{-1} V^\top z$. The goal is to do this calculation while preserving the sparsity of Z and Λ, which requires operating block-wise.

Utilizing the decomposition from equation 4 and recalling our notation that $\tilde{u}$ and $\tilde{\beta}$ denote the mode, we have

$$
\begin{aligned}
\begin{bmatrix} \tilde{u} \\ \tilde{\beta} \end{bmatrix} &= \left(\begin{bmatrix} L_Z & 0 \\ L_{ZX} & L_X \end{bmatrix} \begin{bmatrix} L_Z^\top & L_{ZX}^\top \\ 0 & L_X^\top \end{bmatrix} \right)^{-1} \begin{bmatrix} \Lambda^\top Z^\top & I_Q \\ X^\top & 0 \end{bmatrix} \begin{bmatrix} y \\ 0 \end{bmatrix}, \\
&= \begin{bmatrix} L_Z^{-\top} & -L_Z^{-\top} L_{ZX}^\top L_X^{-\top} \\ 0 & L_X^{-\top} \end{bmatrix} \begin{bmatrix} L_Z^{-1} & 0 \\ -L_X^{-1} L_{ZX} L_Z^{-1} & L_X^{-\top} \end{bmatrix} \begin{bmatrix} \Lambda^\top Z^\top y \\ X^\top y \end{bmatrix}, \\
&= \begin{bmatrix} L_Z^{-\top} & -L_Z^{-\top} L_{ZX}^\top L_X^{-\top} \\ 0 & L_X^{-\top} \end{bmatrix} \begin{bmatrix} L_Z^{-1} \Lambda^\top Z^\top y \\ L_X^{-1} \left(X^\top y - L_{ZX} L_Z^{-1} \Lambda^\top Z^\top y \right) \end{bmatrix} \\
&= \begin{bmatrix} L_Z^{-\top} & -L_Z^{-\top} L_{ZX}^\top L_X^{-\top} \\ 0 & L_X^{-\top} \end{bmatrix} \begin{bmatrix} u' \\ L_X^{-1} \left(X^\top y - L_{ZX} u' \right) \end{bmatrix} && \text{for } u' = L_Z^{-1} \Lambda^\top Z^\top y, \\
&= \begin{bmatrix} L_Z^{-\top} & -L_Z^{-\top} L_{ZX}^\top L_X^{-\top} \\ 0 & L_X^{-\top} \end{bmatrix} \begin{bmatrix} u' \\ \beta' \end{bmatrix} && \text{for } \beta' = L_X^{-1} \left(X^\top y - L_{ZX} u' \right), \\
&= \begin{bmatrix} L_Z^{-\top} \left(u' - L_{ZX}^\top L_X^{-\top} \beta' \right) \\ L_X^{-\top} \beta' \end{bmatrix}.
\end{aligned}
$$

To find the joint mode, one then finds in order $u', \beta', \tilde{\beta}$, and $\tilde{u}$. Every operation involving L_Z and $Z\Lambda$ can be done efficiently thanks to sparsity.

```
> ls()
```

```
 [1] "as.matrix.VarCorr" "commaColumnWrap"   "common_dir"
 [4] "dataDir"           "defaultImgHeight"  "defaultImgWidth"
 [7] "defaultPars"       "display"           "fround"
[10] "functionToString"  "hist.default"      "imgDir"
[13] "makeOrCheckDir"    "oldPars"           "pfround"
[16] "reinstallLibrary"  "reloadLibrary"     "reup"
[19] "sourceDir"         "unloadLibrary"
```

Affiliation:

Vincent Dorie
Department of Humanities and Social Sciences in the Professions
Postdoctoral fellow at the Center for the Promotion of Research Involving Innovative Statistical Methodology
New York University
246 Greene St, Rm 318E
New York, NY 10003, USA E-mail: vjd4@nyu.edu
URL: http://buildmeawebpagesomeday/