Approximately Exact Calculations for Linear Mixed Models

Michael Lavine[1], Andrew Bray[2], and James Hodges[3]

[1]Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003 USA
[2]Reed College, Portland, OR, 97202 USA
[3]Division of Biostatistics, University of Minnesota, Minneapolis, MN, 55455 USA

June 8, 2015

Abstract

This paper is about computations for linear mixed models having two variances, σ_e^2 for residuals and σ_s^2 for random effects, though the ideas can be extended to some linear mixed models having more variances. Researchers are often interested in either the restricted (residual) likelihood $\mathrm{RL}(\sigma_e^2, \sigma_s^2)$ or the joint posterior $\pi(\sigma_e^2, \sigma_s^2 \,|\, y)$ or their logarithms. Both $\log \mathrm{RL}$ and $\log \pi$ can be multimodal and computations often rely on either a general purpose optimization algorithm or MCMC, both of which can fail to find regions where the target function is high. This paper presents an alternative. Letting f stand for either RL or π, we show how to find a box B in the (σ_e^2, σ_s^2) plane such that

1. all local and global maxima of $\log f$ lie within B;
2. $\sup_{(\sigma_e^2,\sigma_s^2)\in B^c} \log f(\sigma_e^2, \sigma_s^2) \le \sup_{(\sigma_e^2,\sigma_s^2)\in B} \log f(\sigma_e^2, \sigma_s^2) - M$ for a prespecified $M > 0$; and
3. $\log f$ can be estimated to within a prespecified tolerance ϵ everywhere in B with no danger of missing regions where $\log f$ is large.

Taken together these conditions imply that the (σ_e^2, σ_s^2) plane can be divided into two parts: B, where we know $\log f$ as accurately as we wish, and B^c, where $\log f$ is small enough to be safely ignored. We provide algorithms to find B and to evaluate $\log f$ as accurately as desired everywhere in B.

1 Introduction

Linear mixed models are an important class of statistical models. Books are written about them (e.g. Bryk and Raudenbush 1992; Verbeke and Molenberghs 2000; Hodges 2013; West et al. 2014), courses are taught about them, and they have many applications. Typical notation, which we adopt, is

$$y = X\beta + Zu + \epsilon \tag{1}$$

where y is a vector of n observations, X is a known $n \times p$ matrix, β is a vector of p unknown coefficients called fixed effects, Z is a known $n \times q$ matrix, u is a vector of q unknown coefficients called random effects, and ϵ is a vector of n errors. The term "mixed" is used when we treat u as a vector of random variables, thus mixing fixed and random effects in the same model. For linear mixed models where u and ϵ are modelled as Normal, researchers are often interested in the restricted likelihood function

$$\mathrm{RL}(\theta) = K|V(\theta)|^{-1/2}|X^tV^{-1}(\theta)X|^{-1/2}\exp\left\{-\frac{1}{2}\left(y^tV^{-1}(\theta)y - \tilde{\beta}^t(\theta)X^tV^{-1}(\theta)X\tilde{\beta}(\theta)\right)\right\} \tag{2}$$

where K is an unimportant constant, θ is a vector of unknown parameters in the covariance matrices of u and ϵ, $V(\theta)$ is the marginal covariance matrix of y implied by the covariance matrices of u and ϵ, and $\tilde{\beta}(\theta)$ is the generalized least-squares estimate of β, given $V(\theta)$. This manuscript deals with the special case in which we adopt the model

$$\epsilon \sim \mathrm{N}(0, \sigma_e^2\Sigma_e) \qquad u \sim \mathrm{N}(0, \sigma_s^2\Sigma_s)$$

where Σ_e and Σ_s are known matrices, often the identity, of the appropriate sizes and $\theta \equiv (\sigma_e^2, \sigma_s^2)$, two unknown variance parameters. The key for this manuscript is that θ contains only those two unknown variances and no others. Examples include random intercept models (including balanced and unbalanced one-way random effect models), additive models with one penalized spline, spatial models with one intrinsic conditional autoregression (ICAR) random effect, dynamic linear models with one system-level variance, and some multiple membership models (e.g. Browne et al. 2001; McCaffrey et al. 2004). Hodges (2013) examines these and other examples and explains the importance of this special case.

Hodges (2013) also unifies and generalizes Reich and Hodges (2008) and Welham and Thompson (2009) to show that in our special case, and a few others, $\log \mathrm{RL}(\sigma_e^2, \sigma_s^2)$ can be expressed as

$$\log \mathrm{RL}(\sigma_e^2, \sigma_s^2) = B - \frac{n_e}{2}\log(\sigma_e^2) - \frac{y^t\Gamma_c\Gamma_c^ty}{2\sigma_e^2} - \frac{1}{2}\sum_{j=1}^{s_z}\left[\log(a_j\sigma_s^2 + \sigma_e^2) + \frac{\hat{v}_j^2}{a_j\sigma_s^2 + \sigma_e^2}\right] \tag{3}$$

where

(1) B is an unimportant known constant;

(2) n_e is n minus the dimension of the space spanned by the columns of $[X|Z]$;

(3) Γ_c is $n \times n_e$ and spans the space orthogonal to $[X|Z]$ (so $y^t\Gamma_c\Gamma_c^ty$ is the residual sum of squares);

(4) s_z is the dimension of the space spanned by the columns of Z not already in the span of the columns of X; and

(5) the $\{a_j\}$ and $\{\hat{v}_j\}$ are known constants whose derivation is in the Appendix. All $a_j > 0$.

Thus the only unknowns are (σ_s^2, σ_e^2) and $\log \mathrm{RL}(\sigma_e^2, \sigma_s^2)$ is a function of just those two arguments.

As Hodges (2013) further observes, if β is given an improper flat prior and σ_e^2 and σ_s^2 are given conjugate priors — say $\sigma_e^2 \sim \mathrm{InvGam}(\alpha_e, \beta_e)$ and $\sigma_s^2 \sim \mathrm{InvGam}(\alpha_s, \beta_s)$ — then

$$-(\alpha_e + 1)\log\sigma_e^2 - \beta_e/\sigma_e^2 - (\alpha_s + 1)\log\sigma_s^2 - \beta_s/\sigma_s^2$$

is added to (3) to yield the log posterior

$$\log \pi(\sigma_e^2, \sigma_s^2 \,|\, y) = B - \frac{n_e + 2\alpha_e + 2}{2} \log(\sigma_e^2) - \frac{y^t \Gamma_c \Gamma_c^t y + 2\beta_e}{2\sigma_e^2} \\ - \frac{2\alpha_s + 2}{2} \log(\sigma_s^2) - \frac{2\beta_s}{2\sigma_s^2} - \frac{1}{2} \sum_{j=1}^{s_z} \left[\log(a_j \sigma_s^2 + \sigma_e^2) + \frac{\hat{v}_j^2}{a_j \sigma_s^2 + \sigma_e^2} \right]. \tag{4}$$

Equations (3) and (4) can both be written as a sum of multiples of logs and inverses of linear combinations $a_j \sigma_s^2 + b_j \sigma_e^2$, as in (5), where the summands with $j = s_z + 1$ and $j = s_z + 2$ are for the terms involving only σ_e^2 and σ_s^2, respectively, and where we have dropped the irrelevant constant B. I.e.,

$$\log f(\sigma_e^2, \sigma_s^2) = -\frac{1}{2} \sum_{j=1}^{s_z+2} \left[c_j \log(a_j \sigma_s^2 + b_j \sigma_e^2) + \frac{d_j}{a_j \sigma_s^2 + b_j \sigma_e^2} \right] \tag{5}$$

where

— for $j = 1, \ldots, s_z$,

$$a_j > 0 \text{ and is derived in the Appendix}$$
$$b_j = 1$$
$$c_j = 1$$
$$d_j = \hat{v}_j^2 \text{ is a known function of } y \text{ derived in the Appendix}$$

— for $j = s_z + 1$,

$$a_j = 0$$
$$b_j = 1$$
$$c_j = \begin{cases} n_e & \text{for } \log \mathrm{RL} \text{ in (3)} \\ n_e + 2\alpha_e + 2 & \text{for the log posterior in (4)} \end{cases}$$
$$d_j = \begin{cases} y^t \Gamma_c \Gamma_c^t y & \text{for } \log \mathrm{RL} \text{ in (3)} \\ y^t \Gamma_c \Gamma_c^t y + 2\beta_e & \text{for the log posterior in (4)} \end{cases}$$

— for $j = s_z + 2$,

$$a_j = 1$$
$$b_j = 0$$
$$c_j = \begin{cases} 0 & \text{for } \log \mathrm{RL} \text{ in (3)} \\ 2\alpha_s + 2 & \text{for the log posterior in (4)} \end{cases}$$
$$d_j = \begin{cases} 0 & \text{for } \log \mathrm{RL} \text{ in (3)} \\ 2\beta_s & \text{for the log posterior in (4).} \end{cases}$$

Our derivation proceeds from (5). We use f to denote the target function generically, either $\mathrm{RL}(\sigma_e^2, \sigma_s^2)$ or $\pi(\sigma_e^2, \sigma_s^2 \,|\, y)$. When the target function is $\mathrm{RL}(\sigma_e^2, \sigma_s^2)$, $c_{s_z+2} = d_{s_z+2} = 0$ and the upper limit of the sum in (5) is effectively $s_z + 1$.

It is known (e.g. Henn and Hodges 2014) that $\log f(\sigma_e^2, \sigma_s^2)$ can have multiple maxima, though the incidence of multiple maxima is unknown. Multi-modal posterior distributions arise readily from conflict

between the likelihood and prior; Liu and Hodges (2003) explores an important such conflict in detail and Wakefield (1998) gives a naturally-occurring example explored further in Henn and Hodges (2014). Multiple maxima in restricted likelihoods have received far less attention and any statement about their incidence would be speculation. To our knowledge, two naturally-occurring cases have have been reported, in Welham and Thompson (2009) and Reiss et al. (2014); Henn and Hodges (2014) report an artificial case and give a recipe for manufacturing examples.

As for numerical optimizers, it is known that existing general purpose algorithms for linear mixed models may fail to find all of the local maxima of $\log f(\sigma_e^2, \sigma_s^2)$, as shown by examples in Hodges (2013), Henn and Hodges (2014), and elsewhere. Mullen (2014) examines 18 optimization functions available in R, tests them on 48 objective functions (admittedly more complicated than $\log f(\sigma_e^2, \sigma_s^2)$) and finds that even the best of them fail in over 10% of the cases. Henn and Hodges (2014) examine conditions under which multiple maxima occur in posterior densities and conclude "...second maxima in posterior distributions therefore may be more common than reports in the literature would suggest." Thus, failure to find local and global maxima may be common, though with available tools it is extremely laborious to determine whether multiple maxima are present and where they are.

Our view is that it is important to find regions where f or $\log f$ is large relative to its maximum regardless of whether those regions contain local maxima. Points with large $\log f(\sigma_e^2, \sigma_s^2)$ are those that describe the data, or possibly prior information, well, at least compared to points with low $\log f(\sigma_e^2, \sigma_s^2)$. If f or $\log f$ is relatively flat and large over a region, it matters little whether the region contains small bumps that are, technically, local maxima. A good analysis should strive to find all points with large $\log f$. Therefore, the purpose of this paper is to introduce an algorithm that will, in finite time, divide the (σ_e^2, σ_s^2) plane into two parts: one where we know $\log f$ is small relative to its maximum and another where we know $\log f$ to within a pre-specified ϵ (hence the term "approximately exact"). The algorithm can also be used to find $(\hat{\sigma}_e^2, \hat{\sigma}_s^2) \equiv \text{argsup}_{\sigma_e^2, \sigma_s^2} \log f(\sigma_e^2, \sigma_s^2)$ (typically either the maximum restricted log likelihood estimate or the maximum *a posteriori* estimate), to within a pre-specified tolerance without fear of missing regions of high f or $\log f$.

The technique relies on the partial derivatives of $\log f(\sigma_e^2, \sigma_s^2)$. Analysis of the partial derivatives allows us to satisfy two desiderata.

D1 For any prespecified constant $M > 0$ we can find a box B, a rectangle in the first quadrant of the (σ_s^2, σ_e^2) plane whose sides are parallel to the axes, such that

$$\text{all local maxima are in } B \qquad \text{and} \qquad \sup_{(\sigma_e^2, \sigma_s^2) \in B^c} \log f(\sigma_e^2, \sigma_s^2) \leq \log f(\hat{\sigma}_e^2, \hat{\sigma}_s^2) - M. \tag{6}$$

In practice we will take M large enough to interpret (6) as meaning that we can restrict attention to B because values of $(\sigma_e^2, \sigma_s^2) \in B^c$ have $\log f(\sigma_e^2, \sigma_s^2)$ too low to be of further interest.

D2 For any box b with sides parallel to the axes we can quickly compute lower and upper bounds (L^b, U^b) satisfying

$$L^b \leq \inf_{(\sigma_e^2, \sigma_s^2) \in b} \log f(\sigma_e^2, \sigma_s^2) \qquad \text{and} \qquad U^b \geq \sup_{(\sigma_e^2, \sigma_s^2) \in b} \log f(\sigma_e^2, \sigma_s^2)$$

and such that $U^b - L^b \to 0$ as b shrinks. Therefore, partitioning the box B from **D1** allows us to know $\log f(\sigma_e^2, \sigma_s^2)$ everywhere in B to within a pre-specified tolerance and also to locate $\text{argsup} \log f$ to within a pre-specified tolerance without fear of missing regions of high $\log f$.

D1 and **D2** allow us to divide the (σ_e^2, σ_s^2) plane into two parts: one where $\log f$ is at least M below its maximum and another where we know $\log f$ to within a pre-specified ϵ. The next section shows how the partial derivatives are used to satisfy **D1** and **D2**: the partial derivatives determine lines in the (σ_e^2, σ_s^2) plane; those lines determine a box B_1 containing all local maxima and a larger box $B \supset B_1$ satisfying **D1**; and those lines also determine upper and lower bounds on $\log f$ within any box b.

2 Satisfying the Desiderata

2.1 Partial Derivatives Determine Lines

The partial derivatives of $\log f(\sigma_e^2, \sigma_s^2)$ can be calculated from (5):

$$\frac{\partial \log f(\sigma_e^2, \sigma_s^2)}{\partial \sigma_s^2} = -\frac{1}{2}\sum_j \left[\frac{a_j c_j}{a_j\sigma_s^2 + b_j\sigma_e^2} - \frac{a_j d_j}{(a_j\sigma_s^2 + b_j\sigma_e^2)^2}\right] = -\frac{1}{2}\sum_j \frac{a_j\left(a_j c_j \sigma_s^2 + b_j c_j \sigma_e^2 - d_j\right)}{(a_j\sigma_s^2 + b_j\sigma_e^2)^2} \tag{7a}$$

and

$$\frac{\partial \log f(\sigma_e^2, \sigma_s^2)}{\partial \sigma_e^2} = -\frac{1}{2}\sum_j \left[\frac{b_j c_j}{a_j\sigma_s^2 + b_j\sigma_e^2} - \frac{b_j d_j}{(a_j\sigma_s^2 + b_j\sigma_e^2)^2}\right] = -\frac{1}{2}\sum_j \frac{b_j\left(a_j c_j \sigma_s^2 + b_j c_j \sigma_e^2 - d_j\right)}{(a_j\sigma_s^2 + b_j\sigma_e^2)^2}. \tag{7b}$$

We work with one term in (7)'s summations at a time; that is, one j at a time. For $j = 1, \ldots, s_z$, the j'th terms in (7) differ by a multiplicative constant $a_j/b_j = a_j$; they have the same sign as each other and the same sign as $(a_j c_j \sigma_s^2 + b_j c_j \sigma_e^2 - d_j) = (a_j\sigma_s^2 + \sigma_e^2 - d_j)$, which determines a line $\sigma_s^2 = d_j/a_j - \sigma_e^2/a_j$ — call it the j'th line — in the first quadrant of the (σ_s^2, σ_e^2) plane. The j'th line has positive intercept d_j/a_j and negative slope $-1/a_j$. For $j = s_z + 1$, $a_j = 0$, so (7a) $= 0$ and (7) determines a vertical line at $\sigma_e^2 = \sigma_e^{2*} \equiv d_{s_z+1}/c_{s_z+1}$. For $j = s_z + 2$, $b_j = 0$, so (7b) $= 0$ and, if the target function is (4), (7) determines a horizontal line at $\sigma_s^2 = \sigma_s^{2*} \equiv d_{s_z+2}/c_{s_z+2}$, while if the target function is (3), $c_{s_z+2} = d_{s_z+2} = 0$; the term for $j = s_z + 2$ effectively vanishes and there is no horizontal line.

For all j, both partial derivatives of the j'th term are nonnegative below or to the left of the j'th line, 0 on the line, and nonpositive above or to the right of the line, as indicated in Figure 1. The j'th term is constant on the j'th line and attains its maximum there.

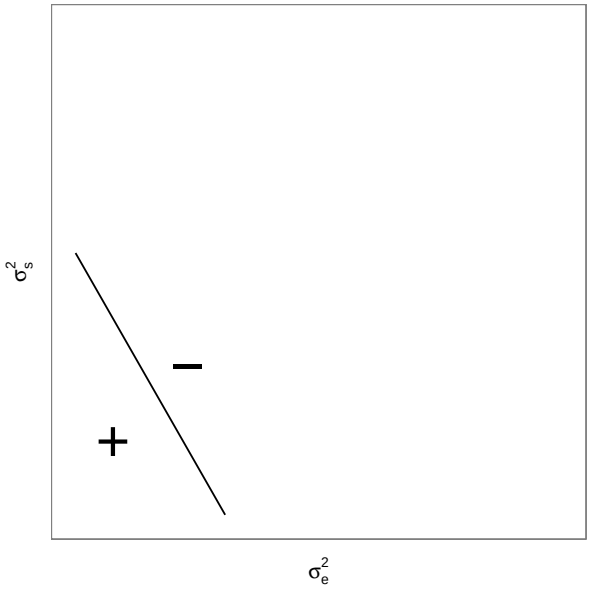


Figure 1: For a single summand, i.e., a fixed j, in (5), the partial derivatives (7a) and (7b) are 0 on the line and nonnegative and nonpositive where indicated by "+" and "-".

2.2 Lines Determine a Bounding Box

Let σ_e^{2M} and σ_s^{2M} be the largest intercepts of the $s_z + 2$ lines on the σ_e^2 and σ_s^2 axes, respectively. I.e.,

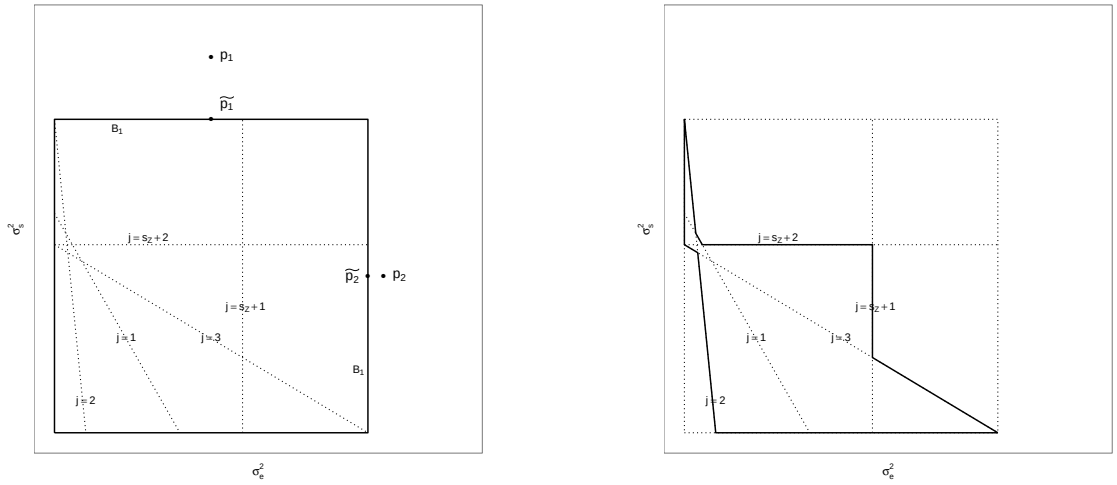
$$\sigma_e^{2M} = \max\left\{\frac{d_1}{c_1}, \cdots, \frac{d_{s_z+1}}{c_{s_z+1}}\right\}$$

$$\sigma_s^{2M} = \begin{cases} \max\left\{\frac{d_1}{a_1}, \cdots, \frac{d_{s_z}}{a_{s_z}}\right\} & \text{if } \log f \text{ is (3)} \\ \max\left\{\frac{d_1}{a_1}, \cdots, \frac{d_{s_z}}{a_{s_z}}, \frac{d_{s_z+2}}{c_{s_z+2}}\right\} & \text{if } \log f \text{ is (4).} \end{cases}$$

Let B_1 be the box whose lower-left and upper-right corners are $(0, 0)$ and $(\sigma_e^{2M}, \sigma_s^{2M})$, respectively. Figure (2a) shows B_1 and five lines — realistic data sets may have more — labelled $j = 1, 2, 3, s_z + 1, s_z + 2$.

Claim *All local maxima of* $\log f$ *lie within* B_1.

Proof Let $p_1 = (\sigma_{e1}^2, \sigma_{s1}^2)$ be a point such that $\sigma_{s1}^2 > \sigma_s^{2M}$ and let $\widetilde{p_1} = (\sigma_{e1}^2, \sigma_s^{2M})$, as illustrated in Figure (2a). For $j = 1, \ldots, s_z, s_z + 2$, the partial derivatives (7b) are negative everywhere between p_1 and $\widetilde{p_1}$; for $j = s_z + 1$, the partial derivative is zero. Therefore, $\log f(\widetilde{p_1}) \geq \log f(p_1)$ and there can be no local maxima of $\log f$ above the line $\sigma_s^2 = \sigma_s^{2M}$. Let $p_2 = (\sigma_{e2}^2, \sigma_{s2}^2)$ be a point such that $\sigma_{e2}^2 > \sigma_e^{2M}$ and let $\widetilde{p_2} = (\sigma_e^{2M}, \sigma_{s2}^2)$. For $j = 1, \ldots, s_z, s_z + 1$, the partial derivatives (7a) are negative everywhere between p_2 and $\widetilde{p_2}$; for $j = s_z + 2$, the partial derivative is zero. Therefore, $\log f(\widetilde{p_2}) \geq \log f(p_2)$ and there can be no local maxima of $\log f$ to the right of the line $\sigma_e^2 = \sigma_e^{2M}$.
QED



(a) A rectangular region B_1 containing all maxima.

(b) A smaller region containing all maxima.

Figure 2: all local and global maxima lie in the regions bounded by the solid dark line.

By the claim, argsup $\log f$ must lie on or inside B_1, so B_1 could be passed to an optimizer such as R's optim or nlminb with potentially better results than using those functions without bounds. However, even with known bounds, general purpose optimizers may still miss argsup $\log f$ and they emphasize single points of highest local $\log f$ while possibly ignoring large regions where $\log f$ is nearly as high (Hill, 1965; Hodges, 2013, Section 18.1.1).

We will next find a box $B \supset B_1$ satisfying desideratum **D1**. First, though, we pause to note that the region outlined in bold in Figure (2b) is a subset of B_1 that must also contain argsup $\log f$ by the same reasoning used to prove the previous claim. But it is not rectangular, hence less convenient than B_1, so we

don't pursue it further.

Claim *For any positive number M, there exist positive numbers $\widetilde{\sigma_e^2} > \sigma_e^{2M}$ and $\widetilde{\sigma_s^2} > \sigma_s^{2M}$ that determine a box $B \supset B_1$ whose lower-left and upper-right corners are $(0,0)$ and $(\widetilde{\sigma_e^2}, \widetilde{\sigma_s^2})$, respectively, such that*

$$\sup_{B^c} \log f(\sigma_e^2, \sigma_s^2) \leq \log f(\hat{\sigma}_e^2, \hat{\sigma}_s^2) - M.$$

I.e., **D1** is satisfied.

Proof Let p be an arbitrary point inside B_1, as illustrated in Figure (3), and define $L \equiv \log f(p)$; we use L as a lower bound on $\log f(\hat{\sigma}_e^2, \hat{\sigma}_s^2)$. Let $\widetilde{q_1} = (0, \widetilde{\sigma_s^2})$, the intercept of B and the σ_s^2 axis; let $\widetilde{q_2} = (\sigma_e^{2*}, \widetilde{\sigma_s^2})$; let $\widetilde{q_3} = (\widetilde{\sigma_e^2}, \sigma_s^{2*})$; let $\widetilde{q_4} = (\widetilde{\sigma_e^2}, 0)$, the intercept of B and the σ_s^2 axis; and let $\log f_j$ denote the j'th term of (5). The proof considers in turn four regions Q_1, Q_2, Q_3, and Q_4 whose union is B^c. Implicitly, B and B^c are functions of $\widetilde{\sigma_e^2}$ and $\widetilde{\sigma_s^2}$. The proof shows that $\widetilde{\sigma_e^2}$ and $\widetilde{\sigma_s^2}$ can be chosen large enough so that **D1** is satisfied on each region.

- $Q_1 : \sigma_e^2 \leq \sigma_e^{2*}$ and $\sigma_s^2 \geq \widetilde{\sigma_s^2}$, as illustrated by q_1 in Figure 3. In Q_1, by inspection of (5) and (7),
  - for $j \neq s_z + 1$,
    $$\frac{\partial \log f_j}{\partial \sigma_e^2} \leq 0 \qquad \text{and} \qquad \frac{\partial \log f_j}{\partial \sigma_s^2} \leq 0$$
    so for any $q \in Q_1$,
    $$\log f_j(q) \leq \log f_j(\widetilde{q_1});$$
  - for $j = s_z + 1$,
    $$\frac{\partial \log f_j}{\partial \sigma_e^2} > 0 \qquad \text{and} \qquad \frac{\partial \log f_j}{\partial \sigma_s^2} = 0$$
    so for any $q \in Q_1$,
    $$\log f_j(q) \leq \log f_j(\widetilde{q_2}).$$

  Therefore,
  $$\log f(\hat{\sigma}_e^2, \hat{\sigma}_s^2) - \log f(q) \geq L - \sum_{j \neq s_z+1} \log f_j(\widetilde{q_1}) - \log f_{s_z+1}(\widetilde{q_2}). \tag{8}$$
  The r.h.s. is larger than M if
  $$\begin{aligned} \sum_{j \neq s_z+1} \log f_j(\widetilde{q_1}) &< L - M - \log f_{s_z+1}(\widetilde{q_2}) \\ &= L - M + \frac{1}{2}\left[c_{s_z+1} \log \sigma_e^{2*} + \frac{d_{s_z+1}}{\sigma_e^{2*}} \right]. \end{aligned} \tag{9}$$
  Examination of (5) shows that for any fixed σ_e^2, in particular for $\sigma_e^2 = 0$, and for $j \neq s_z + 1$, $\lim_{\sigma_s^2 \to \infty} \log f_j(\sigma_e^2, \sigma_s^2) = -\infty$. Thus $\widetilde{\sigma_s^2}$ can be chosen large enough so that the summation on the l.h.s. of (9) is less than the r.h.s. of (9), and **D1** is satisfied.

- $Q_2 : \sigma_e^2 \geq \sigma_e^{2*}$ and $\sigma_s^2 \geq \widetilde{\sigma_s^2}$, as illustrated by q_2 and q_3 in Figure 3. In Q_2, for all j,
  $$\frac{\partial \log f_j}{\partial \sigma_e^2} \leq 0 \qquad \text{and} \qquad \frac{\partial \log f_j}{\partial \sigma_s^2} \leq 0$$
  so for any $q \in Q_2$,
  $$\log f_j(q) \leq \log f_j(\widetilde{q_2}).$$

Therefore, $\log f(\hat{\sigma}_e^2, \hat{\sigma}_s^2) - \log f(q) \geq L - \log f(\widetilde{q_2})$. The latter expression is greater than M iff $\log f(\widetilde{q_2}) < L - M$. But (5) shows that for any fixed σ_e^2, $\lim_{\sigma_s^2 \to \infty} \log f(\sigma_e^2, \sigma_s^2) = -\infty$ and therefore $\widetilde{\sigma_s^2}$ can be chosen large enough so that **D1** is satisfied.

- $Q_3 : \sigma_e^2 \geq \widetilde{\sigma_e^2}$ and $\sigma_s^2 \geq \sigma_s^{2*}$, as illustrated by q_3. In Q_3, for all j,

$$\frac{\partial \log f_j}{\partial \sigma_e^2} \leq 0 \qquad \text{and} \qquad \frac{\partial \log f_j}{\partial \sigma_s^2} \leq 0$$

so for any $q \in Q_3$,

$$\log f_j(q) \leq \log f_j(\widetilde{q_3}).$$

Therefore, $\log f(\hat{\sigma}_e^2, \hat{\sigma}_s^2) - \log f(q) \geq L - \log f(\widetilde{q_3})$. The latter expression is greater than M iff $\log f(\widetilde{q_3}) < L - M$. But for any fixed σ_s^2, $\lim_{\sigma_e^2 \to \infty} \log f(\sigma_e^2, \sigma_s^2) = -\infty$ and therefore $\widetilde{\sigma_e^2}$ can be chosen large enough so that **D1** is satisfied.

- $Q_4 : \sigma_e^2 \geq \widetilde{\sigma_e^2}$ and $\sigma_s^2 \leq \sigma_s^{2*}$, as illustrated by q_4. In Q_4,
  - for $j \neq s_z + 2$,

$$\frac{\partial \log f_j}{\partial \sigma_e^2} < 0 \qquad \text{and} \qquad \frac{\partial \log f_j}{\partial \sigma_s^2} \leq 0$$

so for any $q \in Q_4$,

$$\log f_j(q) \leq \log f_j(\widetilde{q_4});$$

  - for $j = s_z + 2$,

$$\frac{\partial \log f_j}{\partial \sigma_e^2} = 0 \qquad \text{and} \qquad \frac{\partial \log f_j}{\partial \sigma_s^2} \geq 0$$

so for any $q \in Q_4$,

$$\log f_j(q) \leq \log f_j(\widetilde{q_3}).$$

Therefore,

$$\log f(\hat{\sigma}_e^2, \hat{\sigma}_s^2) - \log f(q) \geq L - \sum_{j \neq s_z+2} \log f_j(\widetilde{q_4}) - \log f_{s_z+2}(\widetilde{q_3})$$

which is larger than M if

$$\begin{aligned} \sum_{j \neq s_z+2} \log f_j(\widetilde{q_4}) &< L - M - \log f_{s_z+2}(\widetilde{q_3}) \\ &= L - M + \frac{1}{2}\left[c_{s_z+2} \log \sigma_s^{2*} + \frac{d_{s_z+2}}{\sigma_s^{2*}}\right]. \end{aligned} \tag{10}$$

For any fixed σ_s^2, in particular for $\sigma_s^2 = 0$, and for $j \neq s_z + 2$,
$\lim_{\sigma_e^2 \to \infty} \log f_j(\sigma_e^2, \sigma_s^2) = -\infty$. Thus $\widetilde{\sigma_e^2}$ can be chosen large enough so that the summation on the l.h.s. of (10) is less than the r.h.s. of (10), and **D1** is satisfied.

QED

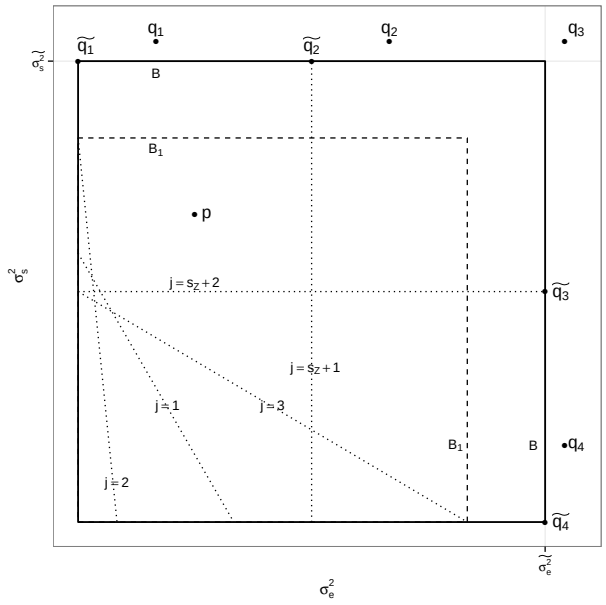


Figure 3: Box B satisfies (6)

2.3 Lines Determine Bounds within Boxes

Next we consider the relationship between the $\log f_j$'s and boxes as depicted in Figure 4, which shows a small box b and three lines. For lines such as $j = 2$ that lie below or to the left of b, the partial derivatives (7a, 7b) are nonpositive, so the maximum and minimum of $\log f_j$ within b are attained at the lower left and upper right corners respectively. The situation is reversed for lines like $j = 3$ that lie above or to the right of b: the partial derivatives are nonnegative so the maximum and minimum are attained at the upper right and lower left corners respectively. For lines like $j = 1$ that pass through b, the minimum is attained at either the upper right or lower left corner while the maximum is attained on the line.

For any box b let $L_j^b = \inf_{p\in b} \log f_j(p)$ and $U_j^b = \sup_{p\in b} \log f_j(p)$. We have just shown that L_j^b and U_j^b are easily computable by evaluating $\log f_j$ at either two points (two corners for lines like $j = 2, 3$) or three points (two corners and one on the line for lines like $j = 1$). Armed with the L_j^b's and U_j^b's we can compute bounds on $\log f(\sigma_e^2, \sigma_s^2)$ within b:

$$L^b \equiv \sum_j L_j^b \le \inf_{(\sigma_s^2,\sigma_e^2)\in b} \log f(\sigma_e^2, \sigma_s^2) \le \sup_{(\sigma_s^2,\sigma_e^2)\in b} \log f(\sigma_e^2, \sigma_s^2) \le \sum_j U_j^b \equiv U^b. \tag{11}$$

Because $\log f$ is continuous, $U^b - L^b \to 0$ as b shrinks in both directions, thus satisfying desideratum **D2**.

3 An Algorithm

Section 2.2 shows how to determine a box B_1 containing all local maxima of $\log f(\sigma_e^2, \sigma_s^2)$ and that we can find a bigger box $B \supseteq B_1$ satisfying **D1**. We are about to present an algorithm that uses Section 2.3 to show that any bounded box B^0 can be partitioned into finitely many smaller boxes $B_1^0, \ldots B_p^0$ such that, for each B_i^0 in the partition, we know for pre-specified constants $M, \epsilon > 0$, either (a) $\sup_{B_i^0} \log f < \sup_{B^0} \log f - M$ or (b) $\sup_{B_i^0} \log f - \inf_{B_i^0} \log f < \epsilon$. If B^0 satisfies **D1** and $B^0 \supseteq B_1$ then we have accomplished our goal of dividing the plane into one region where we know $\log f$ to be low and another where we know $\log f$ to within ϵ. One strategy would be to find B as outlined in Section 2.2 and set $B^0 \equiv B$. However, we have

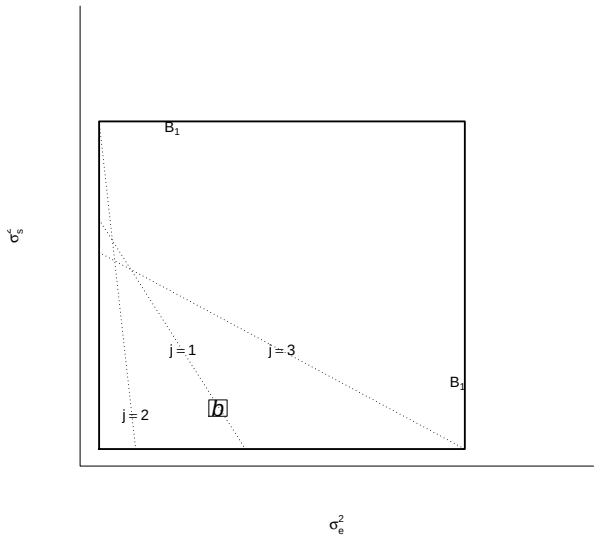


Figure 4: For a box b, the minimum and maximum within b of $\log f_j$ occur at either the lower left corner, the upper right corner, or on the line.

found that B_1 satisfies **D1** in all the examples we have tried and we therefore follow a different strategy. We set $B^0 \equiv B_1$ and run the following algorithm, during which we learn whether B_1 satisfies **D1**. If it does, we're done. If it doesn't, then we would expand B_1 by a factor of 2 in both dimensions; set B^0 to be the expanded box; rerun the algorithm; and continue to expand and rerun until we find a B^0 that does satisfy **D1**. Section 2.2 shows that such a bounded B^0 exists, and thus we are sure to find it in finite time.

Algorithm 1 sketches the basic procedure; details are in an appendix. Before giving examples, we mention some considerations in setting the tuning constants M and ϵ, which have to do with the interpretation of likelihood ratios. We imagine a reference experiment in which a coin is chosen, either fair or two-headed, and tossed. If the coin is tossed twice and lands Heads both times, it produces a likelihood ratio of 4 in favor of the two-headed coin. That's only weak evidence, so we don't see the need to resolve likelihood functions much beyond a factor of four, or loglikelihood functions much beyond a factor of $\log(4) \approx 1.4$. In practice, we use $\epsilon = 1$ as a convenient default. If the coin were tossed 10 times and yielded 10 Heads, the likelihood ratio would be $2^{10} = 1024$. That's strong evidence, so we don't see the need to resolve $\log f$ where it is less than .001 times its maximum. $\log(1000) \approx 7$, so we use $M = 7$ as a convenient default.

4 Examples

Section 4's examples use $\epsilon = 1$ and $M = 7$. Computations were performed on an iMac that was new in 2011, having a 2.7 GHz Intel Core i5 processor and 4GB of memory.

4.1 HMO premiums

Introduction to the Data

Our first example is a traditional linear mixed model previously analyzed in Hodges (1998), Wakefield (1998), Hodges (2013), and Henn and Hodges (2014). Wakefield (1998) reported a bimodal log posterior density for (σ_e^2, σ_s^2). Quoting from Henn and Hodges (2014),

Algorithm 1 Learn areas in B where $\log f$ is high to within ϵ

```
1: function FINDF(B, ε, M)
2:     active ← list containing only B
3:     inactive ← empty list
4:     L ← −∞
5:     while length of active > 0 do
6:         for each element b in active do
7:             L^b, U^b ← get bounds on log f in b                ▷ see section 2.3
8:         end for
9:         tmp ← max(L^b)
10:        L ← max(L, tmp)
11:        for each element b in active do
12:            if U^b − L^b < ε or U^b < L − M then
13:                move b from active to inactive
14:            else
15:                b_1, b_2, b_3, b_4 ← split b into 4 smaller boxes
16:                remove b from active
17:                add b_1, b_2, b_3, b_4 to active
18:            end if
19:        end for
20:    end while
21:    return inactive
22: end function
```

> *. . . the HMO data set describes 341 HMOs [Health Maintenance Organizations] located in 45 states or similar political jurisdictions. Each jurisdiction had between 1 and 31 plans with a median of 5 plans. The data set originally was analysed to assess the cost of moving military retirees and dependents from a Department of Defense health plan to plans serving the US civil service. . . . . Specifically, the model is*
>
> $$y_{ij} = \alpha_i + \epsilon_{ij}$$
> $$\alpha_i = \varrho_0 + \varrho_1 x_{1i} + \varrho_2 x_{2i} + \zeta_i,$$
>
> *where the fixed effects in α_i include an intercept, jurisdiction-average hospital expenses per admission (x_{1i}) and an indicator for plans in New England states (x_{2i}).*

I.e., X is a 341×3 matrix with columns for the intercept and two fixed effects and Z is a 341×45 matrix whose columns are indicators of the 45 jurisdictions. The data are available at `http://www.biostat.umn.edu/~hodges/RPLMBook/Datasets/09_HMO_premiums/Ex9.html`. Because the span of Z's columns contains the span of X's, $s_z = 42$.

A log RL Analysis

For a $\log \mathrm{RL}(\sigma_e^2, \sigma_s^2)$ analysis there are 43 lines, as shown in Figure 5. Running the algorithm on the box B_1 determined by the lines' maximum intercepts on the σ_e^2 and σ_s^2 axes results in the the output displayed in Table 1, which shows the state of the algorithm after iterations 1 through 15: the numbers of active and inactive boxes and the current value of the lower bound L on $\max \log f$. The run finished in less than 10 seconds. We see that boxes are steadily transferred from the active to the inactive list and that L increases monotonically. After 15 iterations there is a total of 9490 boxes, which are displayed in Figure 6.

Figure 6a shows the outlines of the 9490 boxes. The algorithm did not need to divide the boxes with large σ_e^2 or σ_s^2 as finely as those with small σ_e^2 and σ_s^2 because they more readily satisfy either $U^b < L - M$ or $U^b - L^b < \epsilon$, so become inactive. Figure 6b shows the same boxes on a logarithmic scale, shaded by L^b in

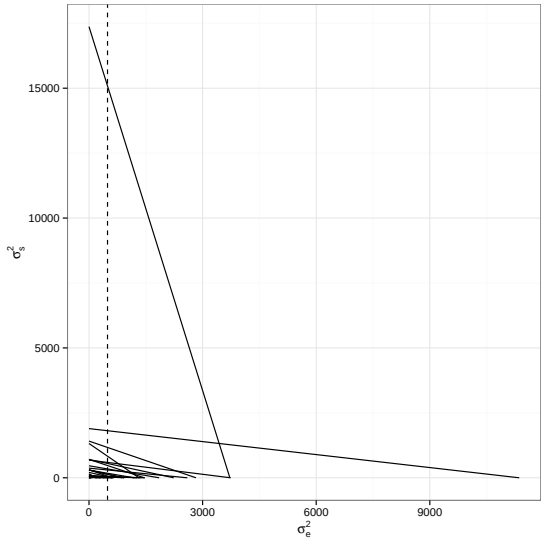
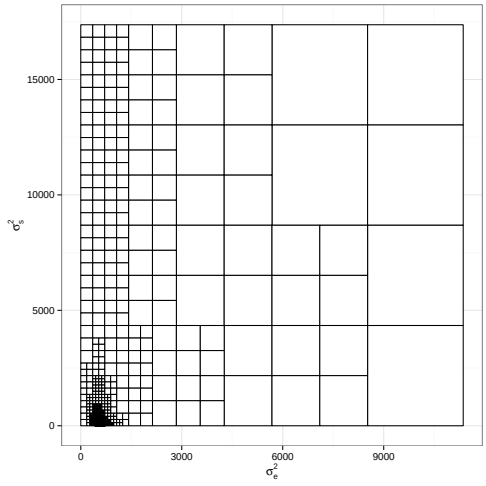
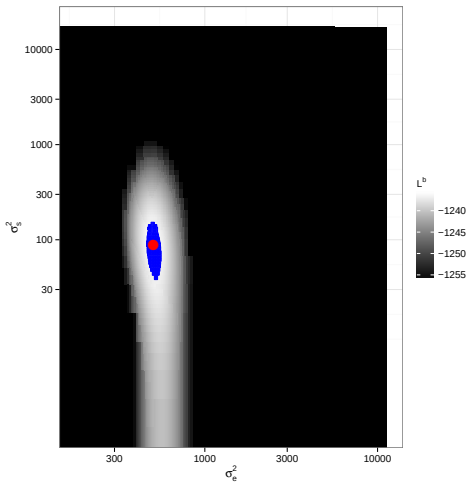


Figure 5: The 43 lines, $j = 1, 2, \ldots, 43$, for the log RL analysis of the HMO data. The line for $j = s_z + 1 = 43$ is dashed.



(a) Locations of the boxes.



(b) Grayscale shows L^b in each box. Red dot shows $b^* \equiv \operatorname{argmax} L^b$. Boxes outlined in blue have $U^b \geq L^{b^*}$. Axes are logarithmic.

Figure 6: The 9490 boxes produced in the log RL analysis of the HMO data. $\mathsf{maxit} = \infty; \epsilon = 1; \delta_e = \delta_s = 0;$ and $M = 7$.

Iteration	n active boxes	n inactive boxes	L
1	4	0	$-\infty$
2	16	0	-1618.84
3	40	6	-1509.24
4	72	28	-1408.12
5	144	64	-1325.99
6	68	191	-1265.01
7	116	230	-1256.35
8	144	310	-1243.51
9	340	369	-1240.05
10	920	479	-1237.87
11	2540	764	-1236.79
12	4380	2209	-1236.20
13	3524	5708	-1235.90
14	344	9146	-1235.90
15	0	9490	-1235.90

Table 1: The state of the algorithm after each of 15 iterations for the HMO data.

each box. The red dot is the lower left corner of the box that maximizes L^b. Boxes with $U^b \geq L \equiv \max L^b$ are outlined in blue; $(\hat{\sigma}_e^2, \hat{\sigma}_s^2)$ must lie within the blue region. For comparison, the standard REML analysis using R's lme function yields $\hat{\sigma}_e^2, \hat{\sigma}_s^2 \approx (495, 99)$ with 95% confidence intervals of $(421, 582)$ and $(39, 248)$.

Figure 6b depicts the same $\log \mathrm{RL}$ as Henn and Hodges (2014)'s Figures 2a (MCMC draws) and 2b ($\log \mathrm{RL}$ contours), but their Figure 2a was produced by MCMC whereas our Figure 6b was produced by direct calculation. Their Figure 2a shows that the MCMC sampler did not sample any values of σ_s^2 less than about 10, whereas our Figure 6b and their Figure 2b show a region of high $\log \mathrm{RL}$ extending down to $\sigma_s^2 = 0$. In fact, $\log \mathrm{RL}(500, 0) \approx -1241.5$, only about 6 log units below $\log \mathrm{RL}(\hat{\sigma}_e^2, \hat{\sigma}_s^2) \approx -1235.7$. Further, about their Figure 2a, Henn and Hodges say, "No change in contour shape indicative of a local maximum could be found in the ... region of $(500, 600) \times (10^{-3}, 1)$, regardless of contour resolution." I.e., they cannot be sure there are no undiscovered points with large $\log \mathrm{RL}$. In contrast, our algorithm guarantees there are no undiscovered points where $\log \mathrm{RL}$ is more than ϵ above L.

A Bayesian Analysis

Hodges (1998), Wakefield (1998), Hodges (2013), and Henn and Hodges (2014) report Bayesian analyses of the HMO data. Here we reproduce the analysis from Hodges (1998) which used inverse Gamma priors for (σ_e^2, σ_s^2) with $\alpha_e = 1$; $\beta_e = 0$; $\alpha_s = 1.1$; and $\beta_s = 0.1$. (We don't defend the prior; we use it so we can compare to Hodges.)

For a Bayesian analysis there are 44 lines, as shown in Figure 7. Figure 7 differs from Figure 5 in that it includes a horizontal line for $j = s_z + 2$ and the position of the vertical line for $j = s_z + 1$ is slightly shifted. Because the $\log \mathrm{RL}$ analysis in Figure 6b shows low values for $\sigma_e^2, \sigma_s^2 > 1000$, we run the Bayesian analysis on the box $B^0 = (0, 1000) \times (0, 1000)$. The algorithm finished in 22 iterations and took a little under 3 hours. Table 2 shows the output and Figure 8 shows the boxes.

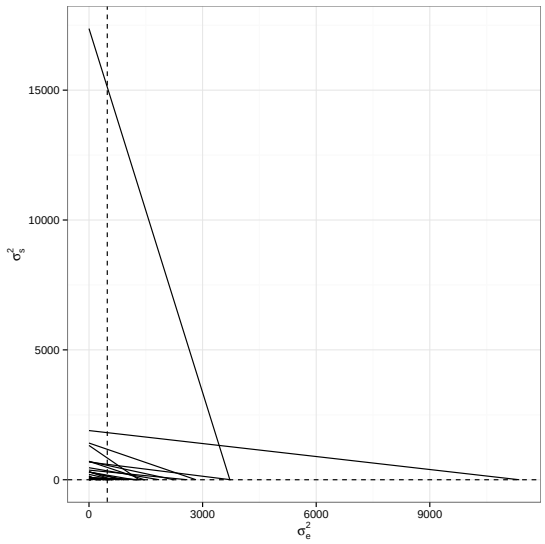
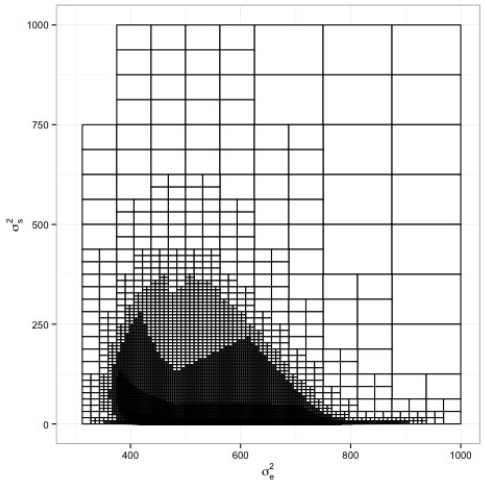


Figure 7: The 44 lines, $j = 1, 2, \ldots, 44$, for the Bayesian analysis of the HMO data. Lines 43 and 44 are dashed.

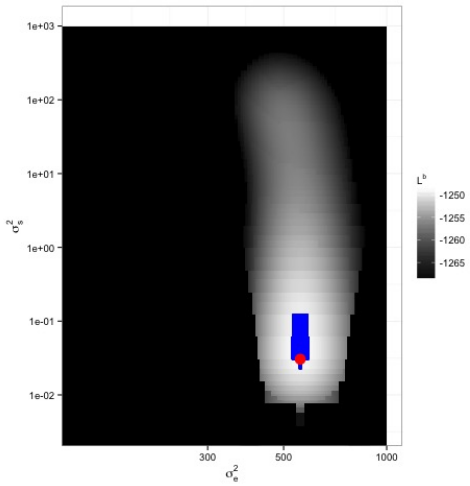
Iteration	n active boxes	n inactive boxes	L
1	4	0	$-\infty$
2	16	0	-1313.42
3	52	3	-1285.05
4	132	22	-1270.76
5	264	88	-1263.98
6	628	195	-1260.47
7	1780	378	-1258.71
8	4188	1111	-1257.70
9	5676	3880	-1256.88
10	5136	8272	-1255.56
11	5528	12026	-1254.20
12	8236	15495	-1252.84
13	12488	20609	-1251.50
14	23852	27134	-1250.26
15	22740	45301	-1249.22
16	43228	57234	-1248.77
17	160208	60410	-1248.77
18	331900	137643	-1248.77
19	575196	325744	-1248.77
20	942860	665225	-1248.77
21	722412	1427482	-1248.77
22	0	2149894	-1248.77

Table 2: The state of the algorithm after each of 22 iterations for the Bayesian analysis of the HMO data.

The Bayesian analysis in Figure 8 can be compared to the log RL analysis in Figure 6. The peak of the log posterior is near $(600, .05)$, very far from the peak of the log RL around $(500, 100)$. The posterior peak is due to the InvGam$(1.1, 0.1)$ prior for σ_s^2, which has a mean of 1, an infinite variance, and a peak at 0.048.



(a) Locations of the boxes.



(b) Grayscale shows L^b in each box. Red dot shows $b^* \equiv \text{argmax}\, L^b$. Boxes outlined in blue have $U^b \geq L^{b^*}$. Axes are logarithmic.

Figure 8: The 2,149,894 boxes in the Bayesian analysis of the HMO data.

The region near $(500, 100)$, though having a lower posterior density, is part of a nearly flat plateau covering a large area. The MCMC draws depicted in Henn and Hodges (2014) Figure 2c show that the plateau has significant posterior mass.

4.2 Global Mean Surface Temperature

We reanalyze a data set in Hodges (2013), global mean surface temperatures (GMST) from 1881 through 2005, depicted in Figure 9. As shown in Ruppert et al. (2003), many splines can be written as linear mixed models. Hodges (2013) fit a piecewise quadratic spline to the GMST data, though a piecewise cubic spline would look similar. Both splines can be formulated as linear mixed models. We follow his lead in fitting a quadratic spline with knots at 1880, 1884, 1888, ..., 2004. X has three columns: $1, \mathsf{year}, \mathsf{year}^2$. Z is 125×30: one row for each year; one column for each knot. Because we fit a quadratic spline, the entries in Z are squares. Σ_e and Σ_s are identity matrices of the appropriate dimension; see Hodges (2013) for details. Following Hodges, we center and scale the year column of X, then compute the year^2 column of X and all the columns of Z from the transformed year, so Z becomes

$$Z = \begin{bmatrix} 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 10.97143 & 10.2521 & \dots & 0.01219048 \\ 11.15505 & 10.42971 & \dots & 0.01904762 \end{bmatrix}$$

Centering and scaling changes only the scale on which σ_s^2 is measured; we do it to more easily compare our result to Hodges'.

The column space of Z shares no dimensions with the column space of X so $s_z = 30$ and, for our $\log \text{RL}$ analysis, there are 31 lines in all, as shown in Figure 10. As usual, we use $B^0 = B_1$, the box determined by the largest intercepts of the 31 lines on the σ_e^2 and σ_s^2 axes. After 25 iterations and about 35 minutes of computing time, all boxes became inactive. Output is in Table 3; boxes are displayed in Figure 11.

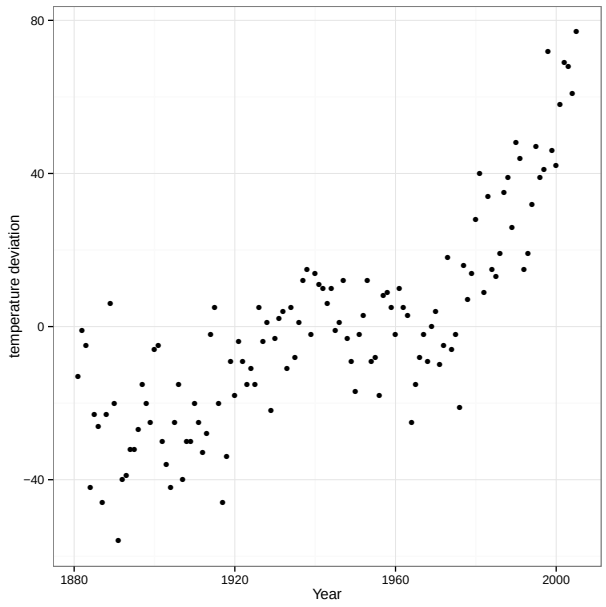


Figure 9: Global mean surface temperature annually from 1881. The y-axis shows deviations from the overall mean in units of .01 degrees C.

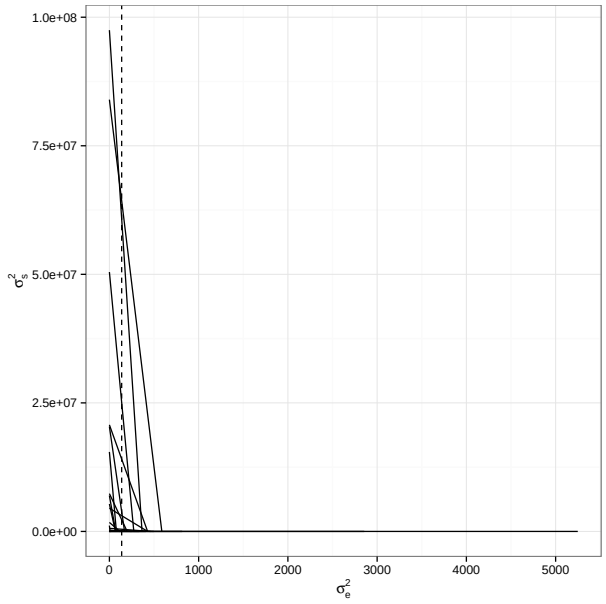
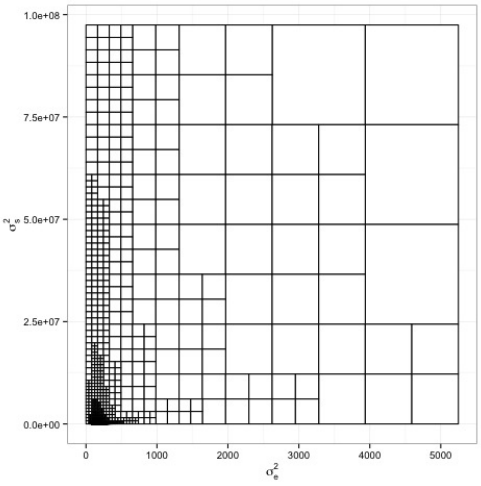
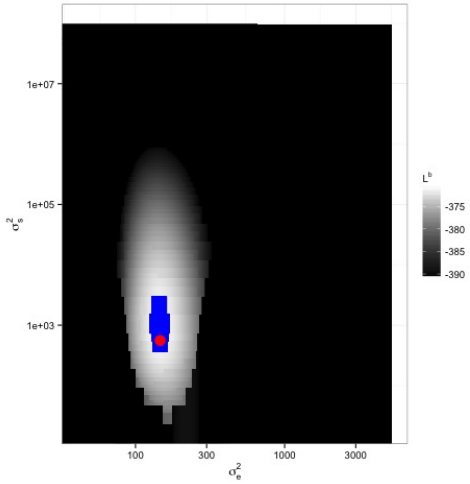


Figure 10: The 31 lines for the log RL analysis of global mean surface temperatures.



(a) Locations of boxes



(b) Grayscale shows L^b in each box. Red dot shows $b^* \equiv \operatorname{argmax} L^b$. Boxes outlined in blue have $U^b \geq L^{b^*}$. Axes are logarithmic.

Figure 11: The 1,474,240 boxes from the analysis of global mean surface temperature.

Iteration	n active boxes	n inactive boxes	L
1	4	0	$-\infty$
2	16	0	-560.82
3	48	4	-519.98
4	84	31	-480.58
5	152	77	-446.67
6	184	183	-419.47
7	176	323	-406.31
8	164	458	-395.36
9	284	551	-390.24
10	496	711	-385.84
11	984	961	-382.45
12	1792	1497	-379.65
13	2880	2569	-377.36
14	4668	4282	-375.37
15	7312	7122	-373.66
16	10336	11850	-372.29
17	10108	19659	-371.21
18	18596	25118	-370.82
19	62488	28092	-370.82
20	124240	59520	-370.70
21	287596	111861	-370.70
22	494292	275884	-370.70
23	646808	608474	-370.70
24	291944	1182296	-370.70
25	0	1474240	-370.70

Table 3: The state of the algorithm after each of 25 iterations for the GMST data.

The figure shows that the algorithm needed to divide boxes near the axes more finely than boxes away from the axes and that high $\log \mathrm{RL}$ is found near $(200, 1000)$. The figure agrees with Figure 15.3 in Hodges (2013).

5 Discussion

This paper explains and illustrates an algorithm that facilitates computation for linear mixed models with two variances. The algorithm finds all regions where either the restricted likelihood function or the joint posterior density of the variances is high and can evaluate the function there to arbitrary accuracy. A natural question to ask is *What about linear mixed models with more than two variances?* A partial answer is given by Hodges (2013) who shows that some models with more than two variances can be re-expressed similarly to (3) but others can't. More complex models that can be re-expressed this way include, but are probably not limited to, models displaying general balance that are also orthogonal designs (all balanced ANOVAs plus other models; Houtman and Speed, 1983), models that are separable in a specific sense (Hodges, 2013, Section 17.1.5), and miscellaneous other models (Hodges, 2013, Section 17.1.5), e.g., a spatial model including random effects for heterogeneity and spatial clustering (an improper conditional autoregressive effect). We have not explored whether the re-expressible models can be analyzed by our algorithm; that's one direction for future work.

Another is to see whether the algorithm can be used to advantage even in non-re-expressible models. If a model has, say, three variances and is now analyzed by, say, MCMC, we can create an MCMC chain that alternates between draws of (σ_e^2, σ_s^2) and draws of the other variance. With the aid of our algorithm we may be able to draw more accurately from $[\sigma_e^2, \sigma_s^2 \,|\, \sigma^2_{\text{other}}]$. More generally, the conditional distribution of (σ_e^2, σ_s^2) given other parameters can now be analyzed more accurately than in the past. We have yet to explore how to exploit that accuracy. A third direction is the posterior $\pi(\sigma_e^2, \sigma_s^2 \,|\, y)$. We can identify a region B^c where the posterior density is low relative to its maximum and it would be of at least mild interest to find an upper bound for the posterior mass of B^c.

As written, our algorithm moves a box b to the inactive list if (a) $U^b < L - M$ or (b) $U^b - L^b < \epsilon$. But one could construct more elaborate rules. One appealing example is to apply criterion (a) if $U^b \le L - \epsilon_2$ and apply criterion (b) if $U^b > L - \epsilon_2$. Other rules are possible, too. We don't elaborate here in order to concentrate on the main ideas.

More generally, our algorithm differs from typical optimization algorithms in that it has a different goal: learning $\log f$ to specified accuracy wherever $\log f$ is high. The algorithm works by exploiting a re-expression of $\log f$ as a sum of simpler, easily analyzed functions. But there may be many other statistically interesting functions that can be so re-expressed. For example, many likelihood functions are products of terms from conditionally independent parts of the data. Posterior densities have the same terms, plus a term from the prior. We have not yet explored whether our algorithm, and more generally the idea of learning $\log f$ to specified accuracy, is useful outside the family of linear mixed models; that's another direction for future work.

In this paper we have taken the point of view that it is important to find all regions where $\log f$ is large without necessarily identifying all local maxima or even the global maximum, even though that point of view is at odds with common statistical estimators that maximize the likelihood, the restricted likelihood, or the posterior density. If two local maxima are close in height it hardly matters which is slightly higher than the other. And, as we said earlier, if there is a high plateau it hardly matters whether there are little bumps on that plateau.

1 Appendix: Derivation of $\{a_j\}$ and $\{\hat{v}_j\}$ in (3), (4), and (5)

Our derivation follows Hodges (2013), which contains more details. There are three steps.

1. **Make the covariance matrices proportional to the identity.** If Σ_e is not the identity matrix, transform the data to $\Sigma_e^{-.5} y$. The transformed data, which we shall still call y, has covariance propor-

tional to the identity. Similarly, if Σ_s is not the identity matrix, re-parameterize the random effects to $\Sigma_s^{-.5}u$. The re-parameterized random effects, which we shall still call u, have covariance proportional to the identity.

2. **If the column spaces of X and Z have a non-trivial intersection, transform them.** Let $s_X = \text{rank}(X)$ and $s_Z = \text{rank}(X|Z) - s_X$. Let Γ_X be an $n \times s_X$ matrix whose columns are an orthonormal basis for the column space of X. Let Γ_Z be an $n \times s_Z$ matrix such that the columns of $[\Gamma_X|\Gamma_Z]$ are an orthonormal basis for the column space of $[X|Z]$. Let Γ_c be an $n \times (n - s_X - s_Z)$ matrix such that the columns of $[\Gamma_X|\Gamma_Z|\Gamma_c]$ are an orthonormal basis for $\mathbb{R}^n$. Define the matrix

$$M = \begin{bmatrix} M_{XX} & M_{XZ} \\ 0 & M_{ZZ} \end{bmatrix}$$

by $[X|Z] = [\Gamma_X|\Gamma_Z]M$ where M_{XX} is $s_X \times p$ and M_{XZ} is $s_Z \times q$. Γ_X and Γ_Z are transformed versions of X and Z that have non-overlapping column spaces.

3. **Re-parameterize and diagonalize.** Let M_{ZZ} have the singular value decomposition $PA^{.5}L^t$. Now the linear mixed model (1) can be written as

$$\begin{aligned} y &= \begin{bmatrix} X|Z \end{bmatrix} \begin{bmatrix} \beta \\ u \end{bmatrix} + \epsilon \\ &= \begin{bmatrix} \Gamma_X|\Gamma_Z \end{bmatrix} M \begin{bmatrix} \beta \\ u \end{bmatrix} + \epsilon \\ &= \begin{bmatrix} \Gamma_X|\Gamma_Z P \end{bmatrix} \begin{bmatrix} \beta^* \\ v \end{bmatrix} + \epsilon \end{aligned}$$

where $\beta^* = M_{XX}\beta + M_{XZ}u$ and $v = A^{.5}L^t u$. β^* contains the re-parametrized fixed effects while v contains the re-parametrized random effects. The corresponding design matrices Γ_X and $\Gamma_Z P$ are orthogonal to each other.

Finally, the $\{a_j\}$ in (3) are the diagonal elements of A, all of which are strictly positive, and the $\{\hat{v}_j\}$ in (3) are given by $\hat{v} = (\hat{v}_1, \ldots, \hat{v}_{s_Z})^t = P^t\Gamma_Z^t y$.

2 Appendix: Details of the algorithm

With **D1** and **D2** the R function findf, sketched below, will evaluate, within a box B^0, $\log f$ to arbitrary accuracy everywhere $\log f$ is large by performing the following tasks: (a) accept as input $\{a_j, b_j, c_j, d_j\}, B^0$ and some tuning constants; (b) create a list of active boxes, initially consisting of just B^0; (c) create a list of inactive boxes, initially empty; and (d) for each active box B_j, find (L^{B_j}, U^{B_j}) and determine whether B_j needs to be subdivided. These notes explain some parts of the function in more detail.

1. A dataframe lines is an input to findf. lines contains the $\{a_j, b_j, c_j, d_j\}$ from Section 2.1. Details for computing them from y, X, Z, Σ_e, and Σ_s are in the Appendix.

2. startbox, or B^0, is an input to findf. As explained at the beginning of Section 3, startbox could be B_1, B, or any other box the user chooses.

3. Constants maxit, M, ϵ, δ_e, and δ_s are inputs to findf.

   (a) Though not mentioned in the main text, maxit is the maximum number (could be ∞) of iterations of findf's loop.

   (b) We will not further analyze regions of the plane where $\log f(\sigma_e^2, \sigma_s^2) < \log f(\hat{\sigma}_e^2, \hat{\sigma}_s^2) - M$. ($M$ could be ∞.)

(c) We will evaluate $\log f$ to within an accuracy of ϵ (could be 0) inside B^0 unless evaluation is stopped by one of the other criteria.

(d) Though not mentioned in the main text, the algorithm can be told not to distinguish values of σ_e^2 separated by less than δ_e (could be 0) nor values of σ_s^2 separated by less than δ_s (could be 0). Separation may be specified in either absolute or relative terms. (I.e. we look at either the difference in σ_e^2 or $\log \sigma_e^2$ (or σ_s^2 or $\log \sigma_s^2$) from one side of the box to the other.)

Setting maxit $= \infty$ and $\delta_e = \delta_s = 0$, as we have done in the examples, implies that B^0 will be partitioned so that either $\log f < \log f(\hat{\sigma}_e^2, \hat{\sigma}_s^2) - M$ or $\log f$ is known to within ϵ, for every B_i^0 in the partition.

4. A box b is a list consisting of
   - upper and lower limits on σ_e^2,
   - upper and lower limits on σ_s^2,
   - upper and lower bounds on $\log f$, U^b and L^b from (11),
   - indicators for whether this box lies above (or to the right of), below (or to the left of), or straddles each of the $s_z + 2$ lines.
5. killfunc is a function, shown below findf, that determines whether a box b should be divided more finely.
6. splitbox is a function that takes a box $b = [\sigma^2_{e\,\mathrm{low}}, \sigma^2_{e\,\mathrm{high}}] \times [\sigma^2_{s\,\mathrm{low}}, \sigma^2_{s\,\mathrm{high}}]$ as input and returns the four boxes created by dividing each side of b at its midpoint.

```
findf <- function(lines, startbox, eps = 0, delE = 0, delS = 0, M = Inf, maxit = 10,
                  ratio = FALSE, lognote = "summary") {


    inactive <- list()  # a list of inactive boxes
    ninact <- 0  # the number of inactive boxes

    active <- list(startbox)  # a list of active boxes
    nact <- 0  # the number of active boxes

    lowbound <- -Inf  # lowerbound on max(log(f))
    iter <- 0  # iteration number

    while (nact > 0 && iter < maxit) {
        # Find the lower bound of each box and the maximum of the lower bounds.  For
        # each active box, either make it inactive or divide it.
        low.act <- max(vapply(X = active, FUN = function(box) {
            box$bounds[1]
        }, FUN.VALUE = 0.1))
        lowbound <- max(lowbound, low.act)
        kill <- vapply(X = active, FUN = killfunc, FUN.VALUE = TRUE, lb = lowbound,
            M = M, eps = eps, delE - delE, delS = delS, ratio = ratio)
            # which boxes become inactive?
        nkill <- sum(kill)
        if (nkill > 0) {
            inactive[(ninact + 1):(ninact + nkill)] <- active[kill]
            # add those boxes to the inactive list
        }
```

```
        ninact <- length(inactive)
        kids <- list()  # subdivisions of active boxes
        nkids <- 0
        for (i in which(!kill)) {  # split active boxes into 4 parts
            kids[(nkids + 1):(nkids + 4)] <- splitbox(active[[i]], lines)
            nkids <- nkids + 4
        }
        active <- kids
        nact <- length(active)

        iter <- iter + 1
        write(c("iteration", iter, "nact", nact, "ninact", ninact, "lowbound",
            lowbound), file = "log.out", ncolumns = 8, append = TRUE)
    }


    tmp <- t(vapply(X = c(active, inactive), FUN = function(box) {
        c(box$lims.sigsqs, box$lims.sigsqe, box$bounds)
    }, FUN.VALUE = c(sigsqs.lo = 0.1, sigsqs.hi = 0.1, sigsqe.lo = 0.1, sigsqe.hi = 0.1,
        rll.lower = 0.1, rll.upper = 0.1)))
    end_time <- Sys.time()
    write(paste(lognote, "runtime: ", round(end_time - start_time, digits = 3)),
          file = "log.out", ncolumns = 1, append = TRUE)
    return(data.frame(tmp))
}

# conditions under which a box becomes inactive
killfunc <- function(box, lb, M, eps, delE, delS, ratio) {
    # lb is a global (within startbox) lower bound on max(log(f));
    #      it changes at each iteration.
    #  M, eps, delE, delS stay constant throughout the iterations.
    cond.low <- box$bounds[2] < lb - M
    cond.eps <- diff(box$bounds) < eps
    cond.E <- ifelse(ratio, diff(log(box$lims.sigsqe)) < delE,
                            diff(box$lims.sigsqe) < delE)
    cond.S <- ifelse(ratio, diff(log(box$lims.sigsqs)) < delS,
                            diff(box$lims.sigsqs) < delS)
    return(cond.low || cond.eps || cond.E || cond.S)  # stop dividing this box?
}
```

References

Browne, W., Goldstein, H., and Rasbash, J. (2001), "Multiple Membership Multiple Classification (MMMC) Models," *Statistical Modeling*, 1, 103–124.

Bryk, A. S. and Raudenbush, S. (1992), *Hierarchical Linear Models: Applications and Data analysis Methods*, Sage, Newbury Park.

Henn, L. and Hodges, J. S. (2014), "Multiple Local Maxima in Restricted Likelihoods and Posterior Distributions for Mixed Linear Models," *International Statistical Review*, 82, 90–105.

Hill, B. (1965), "Inference about variance components in the one-way model," *Journal of the American Statistical Association*, 60, 806–825.

Hodges, J. (1998), "Some algebra and geometry for hierarchical models applied to diagnostics," *jrssb*, 60, 497–536.

Hodges, J. S. (2013), *Richly Parameterized Linear Models: additive, time series, and spatial models using random effects*, CRC Press.

Houtman, A. and Speed, T. (1983), "Balance in designed experiments with orthogonal block structure," *annals*, 11, 1069–1085.

Liu, J. and Hodges, J. (2003), "Posterior bimodality in the balanced one-way random effects model," *jrssb*, 65, 247–255.

McCaffrey, D., Lockwood, J., Koretz, D., Louis, T., and Hamilton, L. (2004), "Models for value-added modeling of teacher effects," *Journal of Behavioral and Educational Statistics*, 29, 67–101.

Mullen, K. M. (2014), "Continuous Global Optimization in R," *Journal of Statistical Software*, 60.

Reich, B. and Hodges, J. (2008), "Identification of the variance components in the general two-variance linear model," *JSPI*, 138, 1592–1604.

Reiss, P., Huang, L., Chen, Y., Huo, L., Tarpey, T., and Mennes, M. (2014), "Massively Parallel nonparametric regression with an application to developmental brain mapping," *jcgs*, 23, 232–248.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge University Press, Cambridge.

Verbeke, G. and Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal Data*, Springer, first edn.

Wakefield, J. (1998), "Comment on *Some Algebra and Geometry for Hieerarchical Models Applied to Diagnostics*," *jrssb*, 60.

Welham, S. and Thompson, R. (2009), "A Note on bimodality in the log-likelihood function for penalized spline mixed models," *Computational Statistics and Data Analysis*, 53, 920–931.

West, B. T., Welch, Kathleen, B., and Galecki, A. T. (2014), *Linear Mixed Models: A Practical Guide Using Statistical Software*, CRC Press, second edn.