Approximately Exact Calculations for Linear Mixed Models

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

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

May 12, 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. 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). 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. Hodges (2013) gives 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. Specifically, $\log \mathrm{RL}(\sigma_e^2, \sigma_s^2)$ is a linear combination of logs and inverses of linear functions $a_j\sigma_s^2 + b_j\sigma_e^2$ of the two unknowns.

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

$$\begin{aligned}\log \pi(\sigma_e^2, \sigma_s^2 \,|\, y) &= B - \frac{n_e + 2\alpha_e + 2}{2} \log(\sigma_e^2) - \frac{y^t \Gamma_c \Gamma_c^t y + 2\beta_e}{2\sigma_e^2} \\ &\quad - \frac{2\alpha_s + 2}{2} \log(\sigma_s^2) - \frac{2\beta_s}{2\sigma_s^2} - \frac{1}{2} \sum_{j=1}^{s_z} \left[\log(a_j \sigma_s^2 + \sigma_e^2) + \frac{\hat{v}_j^2}{a_j \sigma_s^2 + \sigma_e^2} \right] \qquad (4) \\ &= B - \frac{1}{2} \sum_j \left[c_j \log(a_j \sigma_s^2 + b_j \sigma_e^2) + \frac{d_j}{a_j \sigma_s^2 + b_j \sigma_e^2} \right]\end{aligned}$$

for known nonnegative constants $\{a_j, b_j, c_j, d_j\}$ — still a linear combination of logs and inverses of linear functions $a_j \sigma_s^2 + b_j \sigma_e^2$. Our derivation will proceed from (4), though it applies also to (3) and other linear combinations of logs and inverses of linear functions $a_j \sigma_s^2 + b_j \sigma_e^2$. We use f to denote such functions generically.

It is known (e.g. Henn and Hodges 2014) that $\log f(\sigma_e^2, \sigma_s^2)$ can have multiple maxima and that existing general purpose algorithms for linear mixed models may fail to find all of them, 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.

Our point of view is that it is important to find regions where f or $\log f$ is large relative to its maximum regardless of whether the regions contain local maxima. (E.g., 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.) This paper shows how to find $(\hat{\sigma}_e^2, \hat{\sigma}_s^2) \equiv \text{argsup}_{\sigma_e^2, \sigma_s^2} \log f(\sigma_e^2, \sigma_s^2)$ globally (typically either the maximum restricted log likelihood estimate or the maximum *a posteriori* estimate) and to evaluate $\log f(\sigma_e^2, \sigma_s^2)$ everywhere, each to within a prespecified tolerance (hence "approximately exact"), 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

$$\text{all local maxima are in } B \qquad \text{and} \qquad \sup_{(\sigma_e^2, \sigma_s^2) \in B^c} \log f(\sigma_e^2, \sigma_s^2) \le \log f(\hat{\sigma}_e^2, \hat{\sigma}_s^2) - M. \qquad (5)$$

In practice we will take M to be large enough to interpret (5) 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 we can quickly compute lower and upper bounds (L^b, U^b) satisfying

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

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 $\log f$.

The next section shows how the partial derivatives are used to satisfy **D1** and **D2**.

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 (4):

$$\frac{\partial \log f(\sigma_e^2, \sigma_s^2)}{\partial \sigma_s^2} = -\frac{1}{2}\sum_j \left[\frac{a_j c_j}{a_j\sigma_s^2 + b_j\sigma_e^2} - \frac{a_j d_j}{(a_j\sigma_s^2 + b_j\sigma_e^2)^2}\right] = -\frac{1}{2}\sum_j \frac{a_j\left(a_j c_j \sigma_s^2 + b_j c_j \sigma_e^2 - d_j\right)}{(a_j\sigma_s^2 + b_j\sigma_e^2)^2} \tag{6a}$$

and

$$\frac{\partial \log f(\sigma_e^2, \sigma_s^2)}{\partial \sigma_e^2} = -\frac{1}{2}\sum_j \left[\frac{b_j c_j}{a_j\sigma_s^2 + b_j\sigma_e^2} - \frac{b_j d_j}{(a_j\sigma_s^2 + b_j\sigma_e^2)^2}\right] = -\frac{1}{2}\sum_j \frac{b_j\left(a_j c_j \sigma_s^2 + b_j c_j \sigma_e^2 - d_j\right)}{(a_j\sigma_s^2 + b_j\sigma_e^2)^2}. \tag{6b}$$

We work with one term in (6)'s summations at a time; that is, one j at a time. The j'th terms in (6) differ only by a multiplicative constant a_j/b_j; they have the same sign as each other and the same sign as $(a_j c_j \sigma_s^2 + b_j c_j \sigma_e^2 - d_j)$, which determines a line $\sigma_s^2 = d_j/a_j c_j - b_j\sigma_e^2/a_j$ — call it the j'th line — in the first quadrant of the (σ_s^2, σ_e^2) plane. The line has positive intercept and negative slope because all the constants are nonnegative. Both partial derivatives of the j'th term are positive below the j'th line, 0 on the line, and negative above the line, as indicated in Figure 1. The j'th term is constant on the j'th line and attains its maximum there. Both (3) and (4) contain a term, call it the v'th (for vertical) term, with $a_j = 0$. The corresponding line is vertical at $\sigma_e^2 = \sigma_e^{2v} \equiv d_v/b_v c_v$. The v'th term in (6) is negative to the right of the v'th line, positive to the left, and 0 on the line; hence the v'th term in (4) is maximized on the v'th line. (4) also contains a term, call it the h'th (for horizontal) term, with $b_j = 0$. The corresponding line is horizontal at $\sigma_s^2 = \sigma_e^{2h} \equiv d_h/a_h c_h$. The h'th term in (4) is maximized on the h'th line.

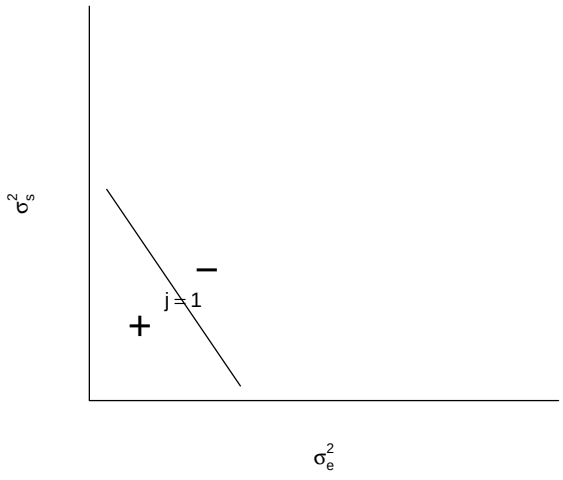
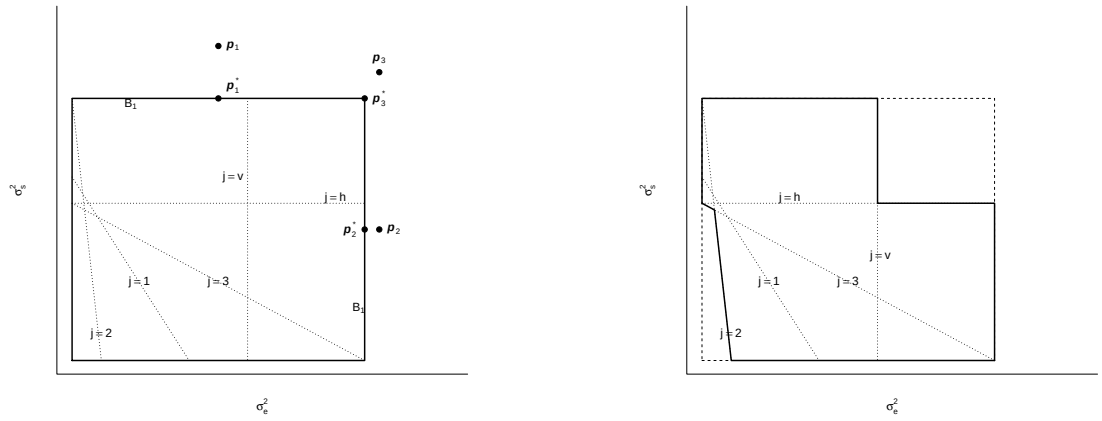


Figure 1: For a single summand, i.e., a fixed j, in (4), the partial derivatives (6a) and (6b) are 0 on the line and positive and negative where indicated by "+" and "-".

2.2 Lines Determine a Bounding Box

Figure (2a) shows five lines — realistic data sets may have more — labelled $j = 1, 2, 3, v, h$. The largest intercept of those lines on the σ_s^2 axis and the largest intercept on the σ_e^2 axis determine a rectangle B_1 whose

outline is marked in bold. At any point above B_1, e.g., p_1, the partial derivatives (6), for all j, are negative or, for $j = v$, zero. Therefore, $\log f(p_1^*) \geq \log f(p_1)$. Similarly, for points like p_2, $\log f(p_2) \geq \log f(p_2^*)$ and, for points like p_3, $\log f(p_3^*) \geq \log f(p_3)$. That is, for points above and/or to the right of B_1, $\log f(\sigma_e^2, \sigma_s^2)$ can be increased by moving down and/or to the left as far as B_1's boundary. Therefore $\operatorname{argsup} \log f$ must lie on or inside B_1 and there can be no maxima of $\log f(\sigma_e^2, \sigma_s^2)$ in B_1^c. 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 $\operatorname{argsup} \log f$ and regions of high $\log f$. This paper presents a computational method guaranteed not to miss $\operatorname{argsup} \log f$ and which evaluates $\log f(\sigma_e^2, \sigma_s^2)$ everywhere to within a specified tolerance inside a box B 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 $\operatorname{argsup} \log f$ for the reasons given above. But it is not rectangular, hence less convenient than B_1, so we don't pursue it further.



(a) A rectangular region 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.

To find B, choose an arbitrary constant $M > 0$ and an arbitrary point inside B_1, say p_4, as illustrated in Figure (3) and define $L \equiv \log f(p_4)$. We will find a box $B \supset B_1$, as shown in Figure (3), outside of which $\log f(\sigma_e^2, \sigma_s^2) \leq L - M$, thus satisfying **D1**. We have drawn a figure without an h line so the reader can see the derivation without it, which would apply to $\log \mathrm{RL}$. If there were an h line, it would influence the derivation for Figure (3)'s q_1 in a manner analogous to how the v line influences the argument for q_2 and q_3.

B is determined by its intercepts σ_e^{2*} and σ_s^{2*} on the σ_e^2 and σ_s^2 axes respectively. Let $q_1^* = (\sigma_e^{2*}, 0)$ be the intercept of B and the σ_e^2 axis. Because each term of the partial derivatives (6) is nonpositive to the right of B_1, for any point q_1 to the right of B, every term in the summation of (4) (with the possible exception of an h term) is larger at q_1^* than at q_1. Therefore, $\log f(\hat{\sigma}_e^2, \hat{\sigma}_s^2) - \log f(q_1) \geq L - \log f(q_1^*)$. The latter expression is greater than M iff $\log f(q_1^*) < L - M$. But examination of (4) shows that for any fixed σ_s^2, in particular for $\sigma_s^2 = 0$, $\lim_{\sigma_e^2 \to \infty} \log f(\sigma_e^2, \sigma_s^2) = -\infty$. Thus by choosing σ_e^{2*} large enough so $\log f(q_1^*) < L - M$, we satisfy (5) for all points with $\sigma_e^2 \geq \sigma_e^{2*}$.

Let $q_2^* = (\sigma_e^{2v}, \sigma_s^{2*})$. For any point q_2 having $\sigma_s^2 \geq \sigma_s^{2*}$ and $\sigma_e^2 \in [\sigma_e^{2v}, \sigma_e^{2*}]$ — i.e., above B and to the right of the v'th line — the partial derivatives are nonpositive so $\log f(\hat{\sigma}_e^2, \hat{\sigma}_s^2) - \log f(q_2) \geq L - \log f(q_2^*)$, which is larger than M if $\log f(q_2^*) < L - M$. But for any fixed σ_e^2, in particular for $\sigma_e^2 = \sigma_e^{2v}$, $\lim_{\sigma_s^2 \to \infty} \log f(\sigma_e^2, \sigma_s^2) = -\infty$. So by choosing σ_s^{2*} large enough we satisfy (5) for all points with $\sigma_s^2 \geq \sigma_s^{2*}$ and $\sigma_e^2 \in [\sigma_e^{2v}, \sigma_e^{2*}]$.

Finally, Let $q_3^* = (0, \sigma_s^{2*})$ be the intercept of B and the σ_s^2 axis. Points such as q_3 which satisfy $\sigma_s^2 \geq \sigma_s^{2*}$ and $\sigma_e^2 \leq \sigma_e^{2v}$ — above B and to the left of the v'th line — are in a region where the partial derivatives (6a, 6b) are negative for $j \in \{1, \ldots, s_z\}$ but nonnegative for $j = v$. Let $\log f_j(\cdot)$ be the j'th term of (3) evaluated

at $(\cdot)$. For $j \geq 1$ we compare $\log f_j(q_3)$ to $\log f_j(q_3^*)$ but we compare $\log f_v(q_3)$ to $\log f_v(q_2^*)$.

$$\log f(\hat{\sigma}_e^2, \hat{\sigma}_s^2) - \log f(q_2) \geq L - \sum_{j=1}^{s_z} \log f_j(q_3^*) - \log f_v(q_2^*) \tag{7}$$

which is larger than M if $\sum_1^{s_z} \log f_j(q_3^*) < L - M - \log f_v(q_2^*) = L - M + n_e(\log(\sigma_e^{2v}) + 1)/2$.

Combining the requirements for points like q_1, q_2, and q_3, we satisfy desideratum **D1** for all points in B^c by choosing σ_e^* large enough to satisfy $\log f(q_1^*) < L - M$ and choosing σ_s^* large enough to satisfy $\log f(q_2^*) < L - M$ and $\sum_1^{s_z} \log f_j(q_3^*) < L - M + n_e(\log(\sigma_e^{2v}) + 1)/2$. In addition, because $\log f$ has no maxima outside of B_1, it also has no maxima outside of B. As already mentioned, if there is an h line, then q_1 would have to be treated like q_2 and q_3.

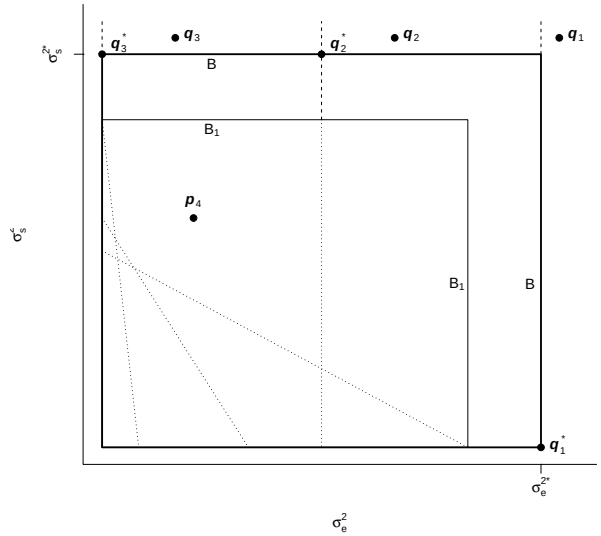


Figure 3: Box B satisfies (5)

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 box b and three lines. (Though we haven't drawn v or h lines, the argument for them is similar to the argument for the lines we have drawn.) For lines such as $j = 2$ that lie below b the partial derivatives of $\log f_j$ are negative, 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 b: the partial derivatives are positive 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:

$$L^b \equiv \sum_j L_j^b \leq \inf_{(\sigma_s^2, \sigma_e^2) \in b} \log f(\sigma_e^2, \sigma_s^2) \leq \sup_{(\sigma_s^2, \sigma_e^2) \in b} \log f(\sigma_e^2, \sigma_s^2) \leq \sum_j U_j^b \equiv U^b. \tag{8}$$

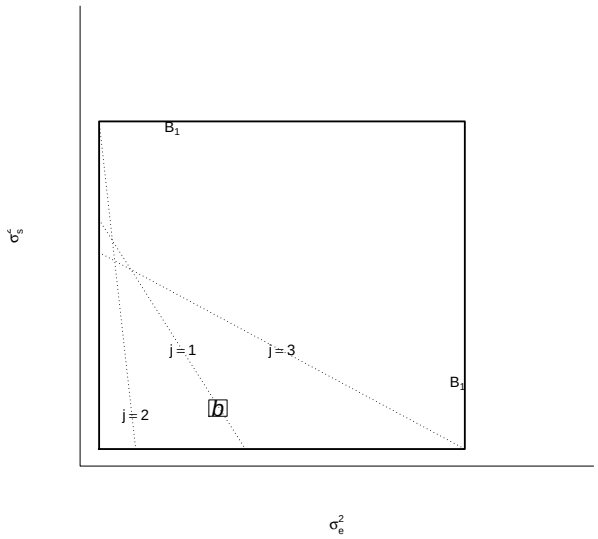


Figure 4: For a box b, the minimum and maximum within b of $\log f_j$ occur at either the lower left corner, the upper right corner, or on the line.

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

3 An Algorithm

With **D1** and **D2** the following algorithm will evaluate $\log f$ to arbitrary accuracy everywhere within a box B.

1. Specify constants maxit, M, ϵ, δ_e, and δ_s.
   (a) maxit is the maximum number (could be ∞) of iterations of the algorithm's loop.
   (b) We will not carefully 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 unless evaluation is stopped by one of the other criteria.
   (d) We will not 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.)

   maxit, M, ϵ, δ_e, and δ_s are constants that control the running of the algorithm. Their meaning may become more clear as we explain the algorithm.
2. With y, X, Z, Σ_e, and Σ_s — see (1) and (2) for definitions — compute $\{a_j, b_j, c_j, d_j\}$ for $j = 1, \ldots, s_z, v, h$.
3. Create two lists of boxes: active and inactive. Both lists are initially empty.
4. Specify a box B as the sole entry of the active list. B could be, but doesn't have to be, a box satisfying desideratum **D1**.

5. Either set $L = -\infty$ or evaluate $\log f(p_i)$ at a few points $p_1, \ldots p_k$ and set $L = \max_i \log f(p_i)$. L is our current lower bound on $\log f(\hat{\sigma}_e^2, \hat{\sigma}_s^2)$.

6. While the active list is not empty or we have not reached maxit iterations

   (a) For each active box b:

      i. if $U^b < L - M$ move b to the inactive list.

      ii. if $U^b - L^b < \epsilon$ move b to the inactive list

      iii. if the horizontal extent of b is less than δ_e, move b to the inactive list

      iv. if the vertical extent of b is less than δ_s, move b to the inactive list

      v. otherwise, b is still active. Divide b into four subboxes. Remove b and add the subboxes to the active list.

   (b) Set $L = \max(L, \max_{b \in \text{active list}} L^b)$. L is our new lower bound on $\log f(\hat{\sigma}_e^2, \hat{\sigma}_s^2)$.

7. Return the active and inactive lists.

4 Examples

4.1 HMO premiums

Introduction to the Data

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

> *... 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. 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 with the settings

$$\mathsf{maxit} = 10; \qquad \epsilon = 5; \qquad \delta_e = \log(10); \qquad \delta_s = \log(10); \qquad M = 5$$

results in the the output displayed in Table 1, which shows the state of the algorithm after iterations 1 through 10: the numbers of active and inactive boxes and the current value of the lower bound L on $\max \log f$. We see that boxes are steadily transferred from the active to the inactive list and that L increases monotonically. After 10 iterations there is a total of 640 boxes, which are displayed in Figure 6.

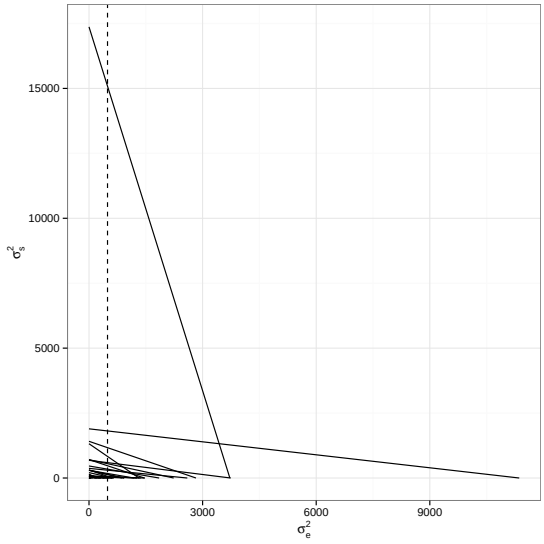
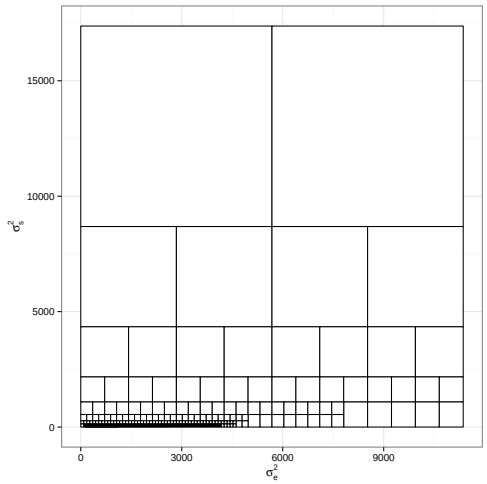
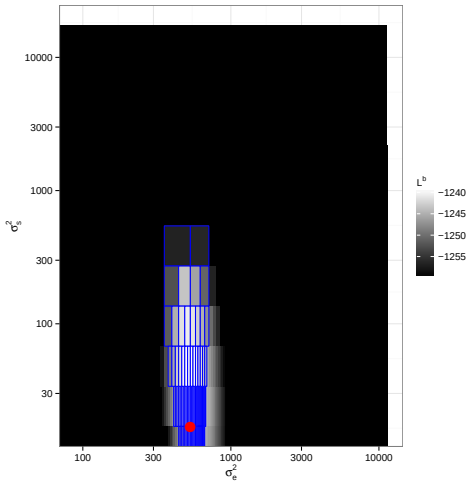


Figure 5: The 43 lines, $j = 1, 2, \ldots, 42, v$, for the log RL analysis of the HMO data. The v line is dashed.



(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 640 boxes produced in the first run for the log RL analysis of the HMO data. $\mathsf{maxit} = 10; \epsilon = 5; \delta_e = \log(10); \delta_s = \log(10);$ and $M = 5$.

Iteration	n active boxes	n inactive boxes	L
1	4	0	$-\infty$
2	8	2	-1618.84
3	16	6	-1509.24
4	32	14	-1408.12
5	44	35	-1325.99
6	56	65	-1265.01
7	104	95	-1256.35
8	184	153	-1243.51
9	224	281	-1240.60
10	180	460	-1239.49

Table 1: The state of the algorithm after each of 10 iterations for the HMO data.

Figure 6a shows the outlines of the 640 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, shaded by L^b in each box. The red dot is the lower left corner of the box that maximizes L^b. Boxes with $U^b \geq \max L^b$ are outlined in blue; $(\hat{\sigma}_e^2, \hat{\sigma}_s^2)$ must lie within the blue region. Figure 6b suggests that we can restrict attention to a box determined by $\sigma_e^2 \in (300, 1000)$ and $\sigma_s^2 \in (0, 1000)$. So we rerun the algorithm on that box and with more stringent control parameters:

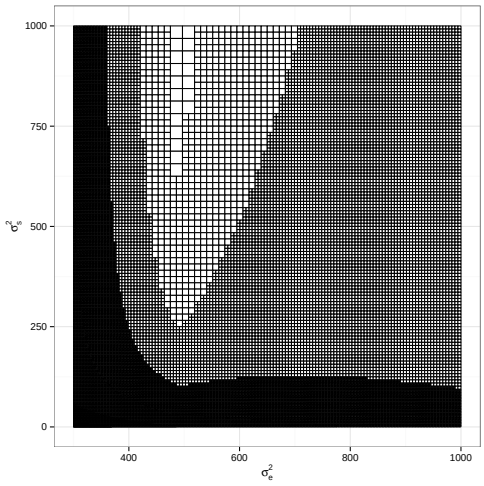
$$\mathsf{maxit} = 15; \qquad \epsilon = 1; \qquad \delta_e = 0; \qquad \delta_s = 0; \qquad M = 10$$

Table 2 shows the output. The algorithm needed 12 iterations to move all boxes to the inactive list. The resulting 37,516 boxes are shown in Figure 7. 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 $(39, 248)$.

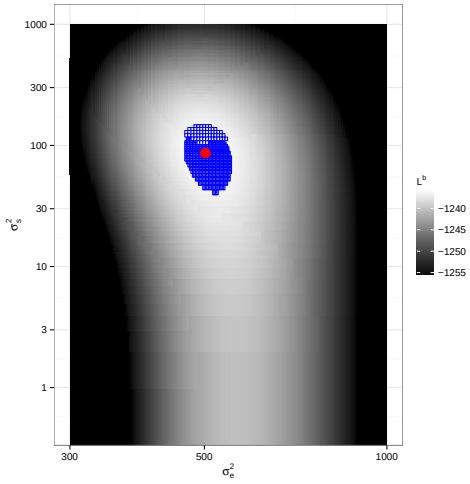
Iteration	n active boxes	n inactive boxes	L
1	4	0	-1320.51
2	16	0	-1276.23
3	64	0	-1252.18
4	256	0	-1243.52
5	1024	0	-1239.61
6	4020	19	-1237.78
7	13128	757	-1236.70
8	14092	10362	-1236.16
9	9444	22093	-1235.88
10	5716	30108	-1235.88
11	2256	35260	-1235.88
12	0	37516	-1235.88

Table 2: The state of the algorithm after each of 12 iterations for the HMO data.

Figure 7b is sufficiently refined that we needn't run the algorithm further. Figure 7b 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 7b 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 7b and their Figure 2b shows that there is 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 $\log \mathrm{RL}(\hat{\sigma}_e^2, \hat{\sigma}_s^2) \approx -1235.5$. 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



(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 7: The 37,516 boxes produced in the second run of the algorithm for the log RL analysis of the HMO data. maxit $= 15$; $\epsilon = 1$; $\delta_e = 0$; $\delta_s = 0$; and $M = 10$.

points with large log RL. In contrast, our algorithm guarantees there are no undiscovered points where log RL is more than ϵ above L.

The HMO log RL analysis illustrates a typical workflow: start with a crudely determined box and coarse settings of the control parameters, then refine the box and the settings as suggested by the output of the first run. Repeat as needed.

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 8. Figure 8 differs from Figure 5 in that it includes a horizontal line and the position of the vertical line is slightly shifted. Because there is only a slight change from the log RL analysis, we began by running the algorithm with the refined control parameters from the log RL analysis. After 15 iterations there were still active boxes, so we ran the algorithm again with the same control parameters except maxit $= 20$. Table 3 shows the output. After 20 iterations there are over 3,000,000 boxes, over 2,000,000 of which are still active.

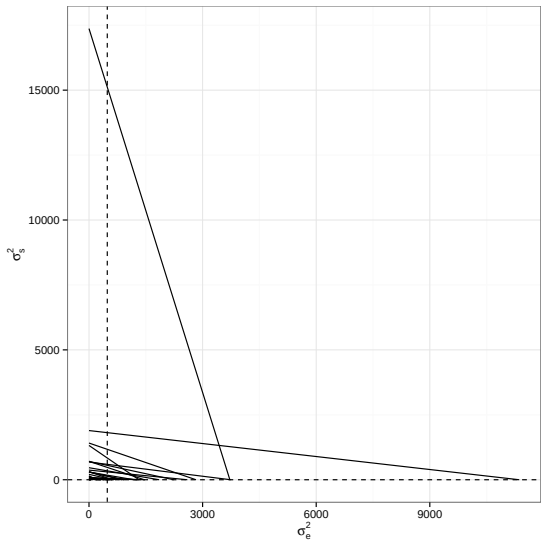
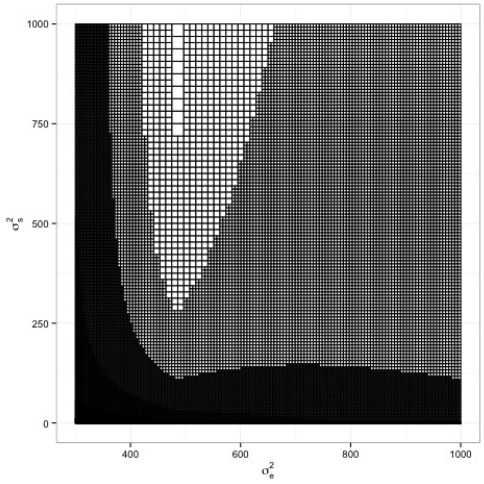


Figure 8: The 44 lines, $j = 1, 2, \ldots, 44, v, h$, for the Bayesian analysis of the HMO data. The v and h lines are dashed.

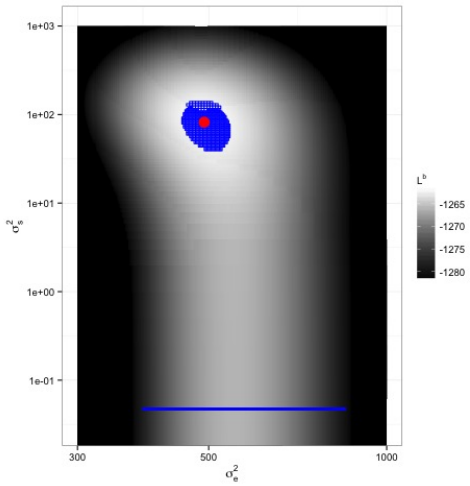
Iteration	n active boxes	n inactive boxes	L
1	4	0	-1350.29
2	16	0	-1303.20
3	64	0	-1279.93
4	256	0	-1270.26
5	1024	0	-1265.89
6	4060	9	-1263.71
7	13740	634	-1262.60
8	14904	10648	-1262.01
9	10180	23007	-1261.71
10	7364	31346	-1261.71
11	6012	37207	-1261.71
12	8192	41171	-1261.71
13	16384	45267	-1261.71
14	32768	53459	-1261.71
15	65536	69843	-1261.71
16	131072	102611	-1261.71
17	262144	168147	-1261.71
18	524288	299219	-1261.71
19	1048576	561363	-1261.71
20	2097152	1085651	-1261.71

Table 3: The state of the algorithm after each of 20 iterations for the Bayesian analysis of the HMO data.

Figure 9 shows the boxes and $\log \pi(\sigma_e^2, \sigma_s^2)$. Figure 9 is similar to Figure 7 except for the addition of the outlined blue boxes near $\sigma_s^2 = 0.04$. In fact, those boxes all have lower boundaries $\sigma_s^2 = 0.04673004$, upper boundaries $\sigma_s^2 = 0.04768372$, $L^b \in (-1270.00, -1266.09)$, $U^b \in (-1252.877, -1248.524)$, and $U^b - L^b \in (17.12299, 18.01643)$, whereas the red dot shows where $\max L^b = -1261.705$ occurs. The ridge in the posterior density near $\sigma_s^2 = 0.04$ is introduced by the InvGam$(1.1, 0.1)$ prior, which has a mean of 1, an



(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 9: The 3,182,803 boxes in the Bayesian analysis of the HMO data. $\mathsf{maxit} = 20; \epsilon = 1; \delta_e = 0; \delta_s = 0;$ and $M = 10$.

infinite variance, and a peak at 0.04761905. The usual analyses of the posterior density are carried out either by MCMC — in which case one hopes that one's Markov chain happens to sample all regions of high density — or by evaluating the posterior on a grid — in which case one hopes one's grid is fine enough to capture all regions of high density. With our analysis there is no need to hope; the algorithm is certain to find all regions of high density.

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 10. As shown in Ruppert et al. (2003), many splines can be written as 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

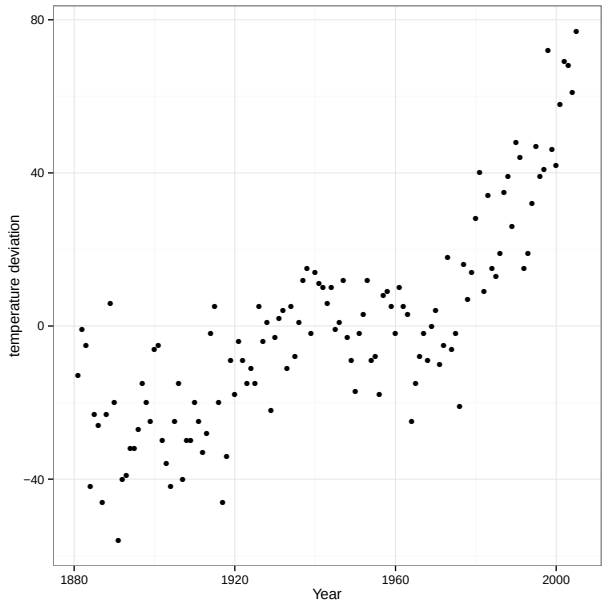


Figure 10: Global mean surface temperature annually from 1881. The y-axis shows deviations from the overall mean in units of .01 degrees C.

is 125×30: one row for each year; one column for each knot.

$$Z = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ 4 & 0 & \dots & 0 \\ 9 & 0 & \dots & 0 \\ 16 & 0 & \dots & 0 \\ 25 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 13924 & 12996 & \dots & 4 \\ 14161 & 13225 & \dots & 9 \\ 14400 & 13456 & \dots & 16 \\ 14641 & 13689 & \dots & 25 \end{bmatrix}; \quad \Sigma_e = \mathbf{1}_{125}; \quad \Sigma_s = \mathbf{1}_{30}$$

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'.

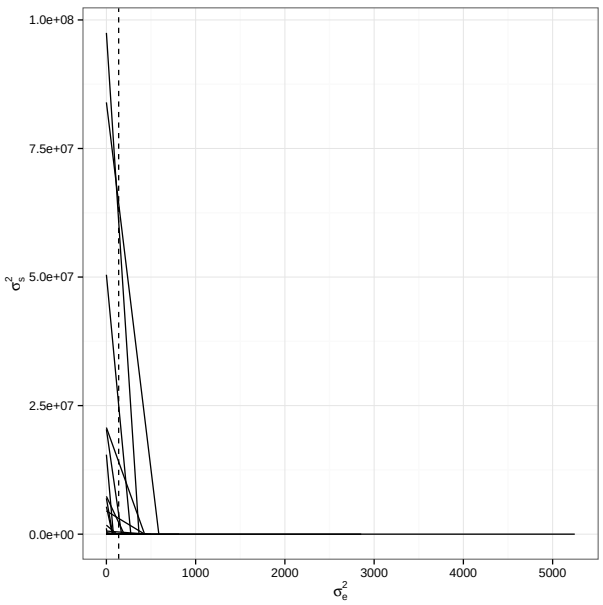


Figure 11: The 31 lines, $j = 1, \ldots, 30, v$, for the log RL analysis of global mean surface temperatures.

The column space of Z shares no dimensions with the column space of X so $s_z = 30$ and, for our log RL analysis, there are 31 lines in all, for $j = 1, \ldots, 31, v$, as shown in Figure 11. Our first run of the algorithm used the box determined by the largest intercepts of the 31 lines on the σ_e^2 and σ_s^2 axes and the control constants

$$\mathsf{maxit} = 15; \qquad \epsilon = 10; \qquad \delta_e = 1; \qquad \delta_s = 1; \qquad M = 10.$$

After 15 iterations there were 3740 boxes still in the active list and 52,973 in the inactive list; they are displayed in Figure 12.

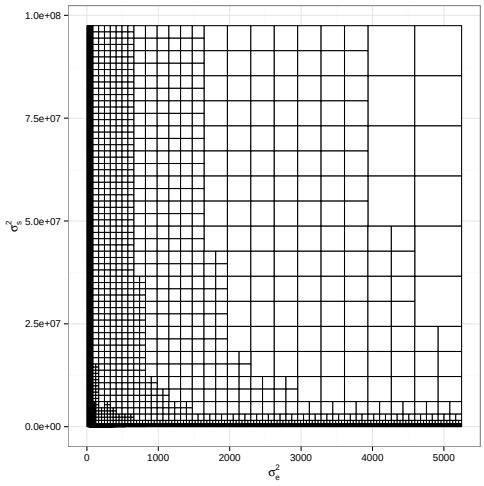
The figure shows that the algorithm needed to divide boxes near the axes more finely than boxes away from the axes and that high log RL is found in only a small region of the plot and therefore suggests rerunning the algorithm with a smaller initial box while imposing more stringent algorithmic parameters. We limited the starting box to $\sigma_e^2 \in (50, 500)$, $\sigma_s^2 \in (0, 10^6)$ and set the control constants to

$$\mathsf{maxit} = 20; \qquad \epsilon = 2; \qquad \delta_e = 0; \qquad \delta_s = 0; \qquad M = 10.$$

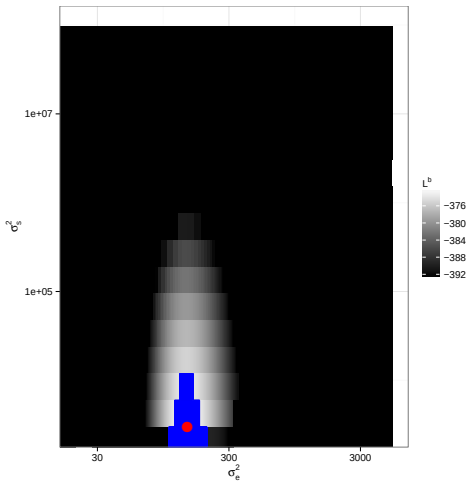
After 20 iterations there were 1,986,000 active and 1,619,095 inactive boxes. Results are in Figure 13. The figure agrees with Figure 15.3 in Hodges (2013) and is detailed enough so no further iterations are needed.

5 Discussion

This paper has explained and illustrated an algorithm, for linear mixed models with two variances, to find all regions where either the restricted likelihood function or the joint posterior density of the variances is high, and to 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 reexpressed similarly to (3) but others can't. More complex models that can be reexpressed 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

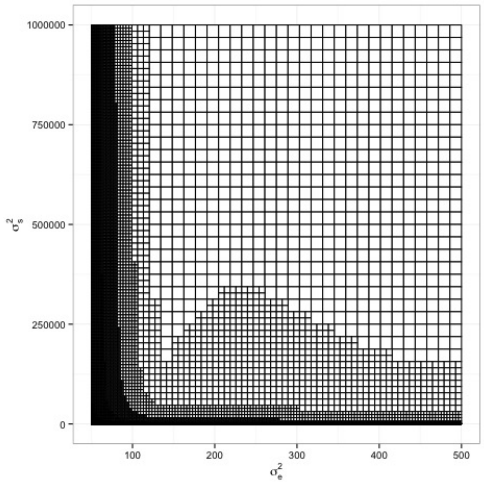


(a) Locations of boxes

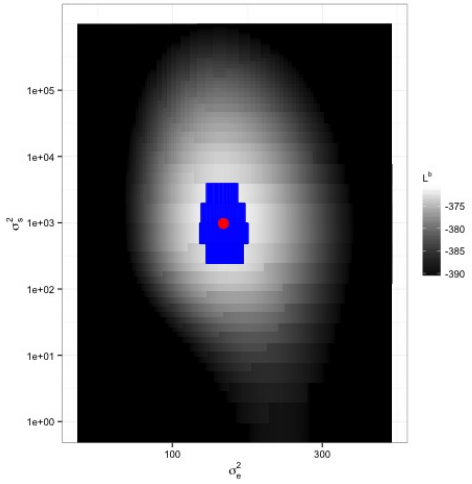


(b) Grayscale shows L^b in each box. Red dot shows $b^* \equiv \operatorname{argmax} L^b$. Boxes outlined in blue have $U^b \geq L^{b^*}$. Axes are logarithmic.

Figure 12: The 56,713 boxes from the first, coarse, analysis of global mean surface temperature.



(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 13: The 3,605,095 boxes from the second analysis of global mean surface temperature.

and spatial clustering (an improper conditional autoregressive effect). We have not explored whether the reexpressible 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-reexpressible 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) $U^b - L^b < \epsilon$ or

(c) either the vertical or horizontal extent of b is sufficiently small.

But one could construct more elaborate rules. One appealing example is to apply criteria (a) and (c) 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.

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.

6 Appendix

Derivation of $\{a_j\}$ and $\{\hat{v}_j\}$ in (3) 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 proportional 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

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

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 $\hat{v} = (\hat{v}_1, \ldots, \hat{v}_{s_Z})^t = P^t \Gamma_Z^t y$.

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. W. (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.

Hodges, J. H. (1998), "Some Algebra and Geometry for Hierarchical Models Applied to Diagnostics," *JRSS B*, 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 of Statistics*, 11, 1069–1085.

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.

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 Hierarchical Models Applied to Diagnostics*," *Journal of the Royal Statistical Society (Series B)*, 60, 497–536.

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.