University of North Carolina at Chapel Hill

The University of North Carolina at Chapel Hill Department of
Biostatistics Technical Report Series

Year 2007

Paper 2

Testing random effects in the linear mixed model using approximate Bayes factors

Benjamin R. Saville*

Amy H. Herring[†]

*University of North Carolina at Chapel Hill, bsaville@bios.unc.edu

[†]University of North Carolina at Chapel Hill, aherring@bios.unc.edu

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

<http://biostats.bepress.com/uncbiostat/art2>

Copyright ©2007 by the authors.

Testing random effects in the linear mixed model using approximate Bayes factors

Benjamin R. Saville and Amy H. Herring

Abstract

Deciding which predictor effects may vary across subjects is a difficult issue. Standard model selection criteria and test procedures are often inappropriate for comparing models with different numbers of random effects, due to constraints on the parameter space of the variance components. Testing on the boundary of the parameter space changes the asymptotic distribution of some classical test statistics and causes problems in approximating Bayes factors. We propose a simple approach for testing random effects in the linear mixed model using Bayes factors. We scale each random effect to the residual variance, and introduce a parameter that controls the relative contribution of each random effect, free of the scale of the data. We integrate out the random effects and the variance components using closed form solutions. The resulting integrals needed to calculate the Bayes factor are low-dimensional integrals lacking variance components, and can be efficiently approximated with Laplace's method. We propose a default prior distribution on the parameter controlling the contribution of each random effect, and conduct simulations to show that our method has good properties for model selection problems. Finally, we illustrate our methods on data from a clinical trial of patients with bipolar disorder and on data from an environmental study of water disinfection by-products and male reproductive outcomes.

Testing random effects in the linear mixed model using approximate Bayes factors

Benjamin R. Saville* and Amy H. Herring

Department of Biostatistics, University of North Carolina at Chapel Hill

CB 7420, Chapel Hill, NC 27599

* *email:* bsaville@bios.unc.edu

SUMMARY: Deciding which predictor effects may vary across subjects is a difficult issue. Standard model selection criteria and test procedures are often inappropriate for comparing models with different numbers of random effects, due to constraints on the parameter space of the variance components. Testing on the boundary of the parameter space changes the asymptotic distribution of some classical test statistics and causes problems in approximating Bayes factors. We propose a simple approach for testing random effects in the linear mixed model using Bayes factors. We scale each random effect to the residual variance, and introduce a parameter that controls the relative contribution of each random effect, free of the scale of the data. We integrate out the random effects and the variance components using closed form solutions. The resulting integrals needed to calculate the Bayes factor are low-dimensional integrals lacking variance components, and can be efficiently approximated with Laplace's method. We propose a default prior distribution on the parameter controlling the contribution of each random effect, and conduct simulations to show that our method has good properties for model selection problems. Finally, we illustrate our methods on data from a clinical trial of patients with bipolar disorder and on data from an environmental study of water disinfection by-products and male reproductive outcomes.

KEY WORDS: Bayes factors; Laplace approximation; mixed model; random effects; variance components

1. Introduction

The linear mixed model with random effects (Laird and Ware, 1982) is a popular method for fitting longitudinal data. In such models it is often of interest to test whether certain random effects should be included in the model. Testing whether a random effect should be included in the model involves the test of whether the variance of that random effect is equal to 0. Because this test lies on the boundary of the parameter space, classical procedures such as the likelihood ratio test can break down asymptotically (Pauler, Wakefield and Kass, 1999; Lin, 1997; Self and Liang, 1987; Stram and Lee, 1994). It has been shown that tests for a single variance component can be carried out using mixtures of chi-square distributions (Self and Liang, 1987; Stram and Lee, 1994). In more complex model comparisons (i.e. testing more than one random effect), distributions of test statistics are more complex and are not easily applied (Pauler et al., 1999; Feng and McCulloch, 1992; Shapiro, 1988). Some alternative frequentist methods include score tests (Lin, 1997; Commenges and Jacqmin-Gadda, 1997; Verbeke and Molenberghs, 2003; Molenberghs and Verbeke, 2007; Zhang and Lin, 2007), Wald tests (Molenberghs and Verbeke, 2007; Silvapulle, 1992), and generalized likelihood ratio tests (Crainiceanu and Ruppert, 2004), but these methods also require modified asymptotic null distributions for tests on the boundary of the parameter space.

Bayesian sampling-based estimation approaches for calculating Bayes factors can also encounter numerical problems on the boundary of the parameter space. Markov chain Monte Carlo (MCMC) methods such as the Gibbs sampler or data augmentation can fail for certain choices of default priors on the random effects (Gilks and Roberts, 1996). Some MCMC methods have been suggested to test variance components (Sinharay and Stern, 2001; Chen and Dunson, 2003; Cai and Dunson, 2006; Kinney and Dunson, 2007), but these methods are generally time consuming to implement, require special software, and rely on subjective choice of hyperparameters which are difficult to elicit. The most widely used approximation

to the Bayes factor is based on the Laplace approximation (Tierney and Kadane, 1986), resulting in the Bayesian information criterion (BIC) (Schwarz, 1978) under certain assumptions. However, the required regularity conditions of the Laplace approximation fail when the parameter lies on the boundary (Pauler et al., 1999; Hsiao, 1997; Erkanli, 1994). Pauler et al. (1999) proposed estimating Bayes factors for model comparison using an importance sampling approach and a boundary Laplace approximation. Their methods are relatively complex and are only applied in the context of simple variance component models.

Because random effects involve a distinct parameter for every individual, linear mixed models can have a very large number of dimensions. This is problematic in calculating Bayes factors, because high dimensional integrals are needed to calculate marginal likelihoods. Generally these integrals are not available in closed form, and one must consider approximations. Numerical integration methods are not efficient nor useful in such high-dimensional integrals (Kuonen, 2003). Monte Carlo integration and importance sampling methods are generally recommended for approximating high-dimensional integrals, but these methods lack accuracy and are computationally demanding. The Laplace and BIC approximations also suffer in performance from high-dimensionality (Kass and Raftery, 1995). In addition, it is not entirely clear how to define the dimensional penalty, or “effective dimension”, in the BIC approximation (Spiegelhalter et al., 2002).

An additional challenge to model selection via Bayes factors lies in the choice of prior distributions. It is well known that Bayes factors are sensitive to the choice of priors (Kass and Raftery, 1995). This is problematic in situations in which one has no prior information on the parameters, and the goal is to choose the best model based on the data. In these situations, it is common to use default priors, which can be chosen based on the data without subjective inputs and that result in good frequentist and Bayesian operating characteristics. However, one must choose these default prior variances with care, because as the prior

variance increases toward infinity the Bayes factor will increasingly favor the null model (Bartlett, 1957). It has been documented that normal priors lead to aberrant behavior in model selection problems, leading Jeffreys (1961) to suggest the Cauchy prior as a heavy-tailed and more robust alternative. This early work by Jeffreys was extended by Zellner and Siow (1980) to develop a robust class of multivariate Cauchy priors for variable selection problems. Zellner's g -prior (Zellner, 1986) has been widely adopted in linear models, and only requires the specification of one hyperparameter. Liang et al. (2005) generalized Zellner's g -prior by implementing a fully Bayes approach using mixtures of g -priors.

In this paper, we propose a simple approach for conducting approximate Bayesian inferences on testing whether to include random effects in the linear mixed model using Bayes factors. Our approach involves a re-parameterization of the linear mixed model, and allows for accurate approximations to the Bayes factor via Laplace's approximation. In Section 2 we introduce our method in the context of a repeated measures ANOVA model, and conduct a simulation to evaluate its performance in testing a subject-specific intercept. In Section 3 we generalize our approach to the linear mixed model, and in Section 4 we illustrate our method using two data examples. We conclude with a discussion in Section 5.

2. Testing a random intercept

2.1 ANOVA model

We start by considering a simple ANOVA model with a random subject effect

$$M_1^{(1)} : Y_{ij} = \mu + \lambda b_i + \varepsilon_{ij}, \quad (1)$$

in which Y_{ij} is the j th response for subject i , μ is an intercept, $b_i \sim N(0, \sigma^2)$ is a scaled random effect multiplied by a parameter $\lambda > 0$, and $\varepsilon_{ij} \sim N(0, \sigma^2)$ is the disturbance term for $i = 1, \dots, n$ and $j = 1, \dots, n_i$. This is an ANOVA model with a random effect variance equal to $\lambda^2 \sigma^2$, in which λ is a parameter controlling the level of within subject correlation.

The utility of this variance component decomposition will later become clear. The notation $M_k^{(a)}$ represents parameterization (a) for model k . We distinguish models parameterized in different ways in order to consider the impact of parameterization on the accuracy of the Laplace approximation to the marginal likelihood. The implied covariance matrix of $\mathbf{y}_i = (Y_{i1}, \dots, Y_{in_i})'$ is $\sigma^2(\mathbf{I}_{n_i} + \lambda^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}')$, in which $\mathbf{I}_{n_i}$ is the $n_i \times n_i$ identity matrix, and $\mathbf{1}_{n_i}$ is a $n_i \times 1$ vector of 1's. It follows that the implied correlation between Y_{ij} and Y_{is} for $j \neq s$ is

$$\rho(Y_{ij}, Y_{is}) = \frac{\lambda^2}{1 + \lambda^2}. \quad (2)$$

Our initial focus is to compare the ANOVA model to a model with no random subject effect,

$$M_0 : Y_{ij} = \mu + \varepsilon_{ij}, \quad (3)$$

in which μ is an overall mean and $\varepsilon_{ij} \sim N(0, \sigma^2)$. We are interested in estimating Bayes factors to determine which model has the largest posterior odds given equal prior odds, given by

$$B_{10}^{(a)} = \frac{p(\mathbf{Y} | M_1^{(a)})}{p(\mathbf{Y} | M_0)}, \quad (4)$$

in which $\mathbf{Y} = (\mathbf{y}'_1, \dots, \mathbf{y}'_n)'$. Estimating the Bayes factor implies deriving estimates of $p(\mathbf{Y} | M_k^{(a)})$, or

$$p(\mathbf{Y} | M_k^{(a)}) = \int p(\mathbf{Y} | \boldsymbol{\theta}_k^{(a)}, M_k^{(a)}) \pi(\boldsymbol{\theta}_k^{(a)} | M_k^{(a)}) d\boldsymbol{\theta}_k^{(a)}, \quad (5)$$

in which $p(\mathbf{Y} | \boldsymbol{\theta}_k^{(a)}, M_k^{(a)})$ is the data likelihood, $\boldsymbol{\theta}_k^{(a)}$ is the vector of model parameters, and $\pi(\boldsymbol{\theta}_k^{(a)} | M_k^{(a)})$ is the prior distribution of $\boldsymbol{\theta}_k^{(a)}$. For clarity, let $M_0^{(a)} = M_0$, i.e. only one parameterization of M_0 will be considered. For $M_1^{(a)}$ and M_0 , the marginal likelihoods are not generally available in closed form for common choices of prior distributions. Let $\boldsymbol{\theta}_1^{(a)} = (\boldsymbol{\zeta}_1'^{(a)}, \mathbf{b}', \sigma^2)'$ and $\boldsymbol{\theta}_0^{(a)} = (\boldsymbol{\zeta}_0'^{(a)}, \sigma^2)'$, such that the vector $\boldsymbol{\zeta}_k^{(a)}$ denotes all parameters other than the random effects $\mathbf{b}$ and residual variance σ^2 . We specify an inverse gamma prior on σ^2 with parameters v, w , in which the mean of σ^2 is $w/(v-1)$ for $v > 1$. By marginalizing

out $\mathbf{b}$ and σ^2 in $M_1^{(1)}$ and σ^2 in M_0 , it can be shown that $(\mathbf{Y}|\mu, \lambda, M_1^{(1)})$ and $(\mathbf{Y}|\mu, M_0)$ follow multivariate t-distributions with the general form

$$p(\mathbf{Y}|\boldsymbol{\zeta}_k^{(a)}, M_k^{(a)}) = \frac{\Gamma\left(\frac{2v+m}{2}\right) \prod_{i=1}^n \left|\frac{w}{v}\boldsymbol{\Sigma}_i\right|^{-1/2}}{(\pi 2v)^{m/2} \Gamma(2v/2)} \times \left\{1 + \frac{1}{2v} \sum_{i=1}^n (\mathbf{y}_i - \boldsymbol{\mu}_i)' \left(\frac{w}{v}\boldsymbol{\Sigma}_i\right)^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i)\right\}^{-\frac{2v+m}{2}}, \quad (6)$$

in which $m = \sum_{i=1}^n n_i$ is the total number of observations. In our ANOVA setup, $\boldsymbol{\mu}_i = \mu \mathbf{1}_{n_i}$ in both M_0 and $M_1^{(1)}$, $\boldsymbol{\Sigma}_i = \mathbf{I}_{n_i}$ in M_0 , and $\boldsymbol{\Sigma}_i = (\mathbf{I}_{n_i} + \lambda^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}') in $M_1^{(1)}$. After specifying a suitable prior on μ , the Laplace method can be used to integrate over (μ, λ) in $M_1^{(1)}$ and μ in M_0 . We then use the resulting marginal likelihood estimates to estimate the Bayes factor $B_{10}^{(1)}$. Note that the vector $\boldsymbol{\mu}_i$ and the matrix $\sigma^2 \boldsymbol{\Sigma}_i$ are the mean and covariance matrix of $\mathbf{y}_i$ marginalized over the random effect b_i . For additional details regarding these multivariate t-distributions, see the Appendices.$

The Laplace approximation is based on a linear Taylor series approximation of $\tilde{l}(\boldsymbol{\zeta}_k^{(a)}) = \log\{p(\mathbf{Y}|\boldsymbol{\zeta}_k^{(a)}, M_k^{(a)})\pi(\boldsymbol{\zeta}_k^{(a)}|M_k^{(a)})\}$. The marginal likelihood $p(\mathbf{Y}|M_k^{(a)})$ for model k and parameterization (a) is estimated by

$$\hat{p}(\mathbf{Y}|M_k^{(a)}) = (2\pi)^{d/2} |\tilde{\boldsymbol{\Sigma}}_k^{(a)}|^{1/2} p(\mathbf{Y}|\tilde{\boldsymbol{\zeta}}_k^{(a)}, M_k^{(a)}) \pi(\tilde{\boldsymbol{\zeta}}_k^{(a)}|M_k^{(a)}), \quad (7)$$

in which $\tilde{\boldsymbol{\Sigma}}_k^{(a)}$ is the Hessian matrix of $\tilde{l}(\boldsymbol{\zeta}_k^{(a)})$ evaluated at the posterior mode $\tilde{\boldsymbol{\zeta}}_k^{(a)}$. Because the Laplace approximation is based on a linear Taylor series approximation, it requires certain regularity conditions. When the posterior mode lies on the boundary of the parameter space these regularity conditions fail. The Laplace method can perform poorly even if the mode is close to the boundary of the parameter space; hence estimating the marginal likelihood in $M_1^{(1)}$ via Laplace can be problematic because of the restricted parameter space of $\lambda > 0$. If the posterior mode $\tilde{\lambda}$ is close to 0, this can cause problems with the accuracy of the approximation. Hence we consider an alternate parameterization of equation (1),

$$M_1^{(2)} : Y_{ij} = \mu + e^\phi b_i + \varepsilon_{ij}, \quad (8)$$

in which $\phi = \log(\lambda)$. Note the parameter space of ϕ is unrestricted, ensuring that the posterior mode falls within the boundaries of the parameter space. Because the posterior mode of ϕ will not violate the regularity conditions of the Laplace approximation, the estimated marginal likelihoods based on $M_1^{(2)}$ may be more accurate than those based on $M_1^{(1)}$. Following the steps outlined previously, it can be shown that $(\mathbf{Y}|\mu, \phi, M_1^{(2)})$ follows a multivariate t-distribution with density (6), with $\boldsymbol{\mu}_i = \mu \mathbf{1}_{n_i}$ and $\boldsymbol{\Sigma}_i = (\mathbf{I}_{n_i} + e^{2\phi} \mathbf{1}_{n_i} \mathbf{1}_{n_i}')$. We use the Laplace approximation to integrate over (μ, ϕ) , and use the resulting estimate of the marginal likelihood to estimate the Bayes factors $B_{10}^{(2)}$.

2.2 Prior choice

It is well understood that a Bayes factor is sensitive to the choice of prior distributions (Kass and Raftery, 1995). As the prior variance of the random effect increases toward infinity, the Bayes factor will increasingly favor M_0 over the random effects model. It is therefore of interest to suggest default priors that yield robust tests with respect to model selection. In our model, we have introduced a parameter λ (or e^ϕ) that controls the contribution of the random effect, free of the scale of the data. We propose default priors of $\lambda \sim \log N(\kappa = 0, \tau = 1)$ and $\phi \sim N(0, 1)$, in which κ and τ denote the mean and variance of the log-normal distribution on the log scale. The priors on λ and ϕ are “equivalent” priors, meaning they lead to the same marginal likelihood. Any differences in the estimated marginal likelihoods between $M_1^{(1)}$ and $M_1^{(2)}$ should be a result of differences in the accuracy of the Laplace approximation under different parameterizations.

In choosing a prior distribution for $\lambda(> 0)$, we want to avoid a prior that is concentrated around the null value of 0. Given that the the random effects are scaled to the residual error, a $\log N(0, 1)$ prior on λ is a reasonable default prior for model selection. After marginalizing out σ^2 , this log normal prior is heavy-tailed and covers most reasonable mean values of the parameter.

2.3 Simulation study

We conducted a simulation study to evaluate the performance of parameterization $M_1^{(2)}$ and $M_1^{(1)}$ in correctly identifying models with or without random effects. We simulated 100 data sets based on parameterization (1), with $n = 25, 50, 100, 500, 1000, 5000$, $n_i = 3$ and $\sigma^2 = 1$. The parameter λ was varied to allow different degrees of correlation in the simulated data, or correlations of 0, 0.14, 0.33, 0.5, 0.69. In order to implement the Laplace approximation, we estimated the posterior mode using an algorithm by Nelder and Mead (1965). We used prior distributions $\mu \sim N(0, 1)$ and $\sigma^2 \sim \text{InvGam}(1, 1)$. Estimates of the Bayes factors $B_{10}^{(1)}$ and $B_{10}^{(2)}$ were calculated for each data set for a given correlation, and were interpreted according to the scale given by Wasserman (2000) and Jeffreys (1961). Table 1 includes the percent of times that the estimated Bayes factors fell into the respective categories, indicating weak, moderate, or strong evidence in favoring a given model.

[Table 1 about here.]

Both parameterizations performed well in favoring the correct model, but accuracy depended on both the sample size and the simulated correlation. In general, as ρ increased our method increasingly favored $M_1^{(a)}$ over the null model. As the sample size increased, our method more accurately detected the absence of a random slope for $\rho = 0$, and more accurately detected the presence of a random slope for $\rho > 0$. For small sample sizes, we observed reasonably good performance with zero or moderate correlation ($\rho = 0, \rho \geq 0.33$). However, larger sample sizes were needed in order to detect smaller correlations close to the boundary. Figure 1 shows box plots of $\log \hat{B}_{10}^{(1)}$ for $\rho = 0, 0.33$. The dotted black line represents a log Bayes factor of 0 (i.e. Bayes factor equal to 1). As n goes toward infinity the estimated log Bayes factor $B_{10}^{(1)}$ goes to infinity for $\rho = 0.33$, and goes to negative infinity for $\rho = 0$, showing that our method increasingly favors the correct model as n increases.

[Figure 1 about here.]

The estimated Bayes factors comparing $M_1^{(a)}$ to M_0 were very similar across parameterizations, even close to the boundary. We also considered the use of numerical integration to more effectively compare the parameterizations. For finite integrals of low dimension, adaptive numerical integration is an accurate and efficient method for calculating integrals. We employed transformations on the parameters (λ, ϕ, μ) to map the infinite integral in (5) to a finite integral, and implemented Genz' (1991) adaptive numerical integration routine for sample sizes $n = 25, 50$. We did not find either parameterization to outperform the other. In fact, the Laplace approximations from the two parameterizations were so close that it was difficult to compute a numerical integration approximation with enough precision to distinguish between the two parameterizations. Hence, given the similarities between the two parameterizations, it is fairly evident that the boundary issue of λ is not a major problem with the Laplace approximation in this model.

3. Testing a random slope

3.1 Linear mixed model

We generalize our approach for testing random effects by considering a linear mixed model (Laird and Ware, 1982) of the form

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad (9)$$

in which $\mathbf{y}_i = (Y_{i1}, \dots, Y_{in_i})'$ is a $n_i \times 1$ vector of responses, $\mathbf{X}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{ip})$ is a $n_i \times p$ design matrix, $\mathbf{Z}_i = (\mathbf{z}_{i1}, \dots, \mathbf{z}_{iq})$ is a $n_i \times q$ design matrix, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ is a $p \times 1$ vector of parameters, and $\mathbf{b}_i = (b_{i1}, \dots, b_{iq})'$ is a $q \times 1$ vector of random coefficients. The matrix $\mathbf{Z}_i$ is usually considered to be a subset of $\mathbf{X}_i$. It is assumed that $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \mathbf{R})$ is independent of $\mathbf{b}_i \sim N(\mathbf{0}, \boldsymbol{\psi})$, in which $\boldsymbol{\psi}$ is the $q \times q$ covariance matrix of random effects. A popular choice for $\mathbf{R}$ is $\sigma^2 \mathbf{I}$, which assumes the observations are independent within a subject given the random coefficients.

We choose $b_{ih} \sim N(0, \sigma^2)$, and introduce a parameter λ_h that controls the relative contribution of the h th random effect of subject i . Let $M_k^{(a)}$ refer to model k and parameterization a . Similar to the approach of Chen and Dunson (2003) (but without the assumption of $\mathbf{b}_{0,i} \sim N(\mathbf{0}, \mathbf{I})$), our reparameterized model takes the form

$$M_0^{(1)} : \mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_{0,i} \boldsymbol{\Lambda}_0^{(1)} \boldsymbol{\Gamma}_0 \mathbf{b}_{0,i} + \boldsymbol{\varepsilon}_i, \quad (10)$$

in which $\mathbf{Z}_{0,i} = (\mathbf{z}_{i1}, \dots, \mathbf{z}_{iq})$, $\mathbf{b}_{0,i} = (b_{i1}, \dots, b_{iq})'$, $\boldsymbol{\Lambda}_0^{(1)}$ is a positive diagonal matrix with diagonal elements $\boldsymbol{\lambda}_0^{(1)} = (\lambda_1, \dots, \lambda_q)'$, and $\boldsymbol{\lambda}_0^{(1)} \sim \log N(\mathbf{0}_q, \mathbf{I}_q)$ as an extension of the “default” prior suggested in 2.2. Let $\boldsymbol{\Gamma}_0$ be a lower triangular matrix with $\mathbf{1}_q$ along the diagonal, and lower off-diagonal elements $\boldsymbol{\gamma}_0$ which induce correlation between the respective random effects.

Our focus is to test whether to include an additional random effect $b_{i(q+1)}$. Let $\mathbf{Z}_{1,i}$, $\boldsymbol{\Lambda}_1^{(1)}$, $\boldsymbol{\Gamma}_1$, and $\mathbf{b}_{1,i}$ be equal to their counterparts from (10), but including the elements corresponding to the additional random effect $b_{i(q+1)}$. The full model including the additional random effect takes the form

$$M_1^{(1)} : \mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_{1,i} \boldsymbol{\Lambda}_1^{(1)} \boldsymbol{\Gamma}_1 \mathbf{b}_{1,i} + \boldsymbol{\varepsilon}_i, \quad (11)$$

in which $\mathbf{Z}_{1,i} = (\mathbf{z}_{i1}, \dots, \mathbf{z}_{i(q+1)})$, $\mathbf{b}_{1,i} = (b_{i1}, \dots, b_{i(q+1)})'$, $\boldsymbol{\Lambda}_1^{(1)}$ is a positive diagonal matrix with diagonal elements $\boldsymbol{\lambda}_1^{(1)} = (\lambda_1, \dots, \lambda_{q+1})'$ and $\boldsymbol{\lambda}_1^{(1)} \sim \log N(\mathbf{0}_{q+1}, \mathbf{I}_{q+1})$, and $\boldsymbol{\Gamma}_1$ is a lower triangular matrix with $\mathbf{1}_{q+1}$ along the diagonal and lower off-diagonal elements $\boldsymbol{\gamma}_1$.

As demonstrated with the ANOVA model, we also consider an alternate parameterization of (10) and (11), by setting $\boldsymbol{\lambda}_k^{(2)} = \boldsymbol{\phi}_k = \log \boldsymbol{\lambda}_k^{(1)}$, with $\boldsymbol{\lambda}_0^{(2)} \sim N(\mathbf{0}_q, \mathbf{I}_q)$ and $\boldsymbol{\lambda}_1^{(2)} \sim N(\mathbf{0}_{q+1}, \mathbf{I}_{q+1})$. We define $\boldsymbol{\Lambda}_0^{(2)}$ as a diagonal matrix with diagonal elements $e^{\boldsymbol{\lambda}_0^{(2)}} = (e^{\phi_1}, \dots, e^{\phi_q})$, and $\boldsymbol{\Lambda}_1^{(2)}$ as a diagonal matrix with diagonal elements $e^{\boldsymbol{\lambda}_1^{(2)}} = (e^{\phi_1}, \dots, e^{\phi_{q+1}})$. Let $M_0^{(2)}$ denote the reduced model and $M_1^{(2)}$ denote the full model under this parameterization.

3.2 Approximating the marginal likelihoods

In order to implement the Laplace approximation, we first marginalize out $\mathbf{b}$ and σ^2 . Let $\sigma^2 \sim \text{InvGam}(v, w)$. It can be shown that the marginal distribution $p(\mathbf{Y}|\boldsymbol{\beta}, \boldsymbol{\lambda}_k^{(a)}, \boldsymbol{\gamma}_k, M_k)$ follows a multivariate t-distribution with density (6), in which $\boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_i = (\mathbf{I}_{n_i} + \mathbf{Z}_{k,i}\boldsymbol{\Lambda}_k^{(a)}\boldsymbol{\Gamma}_k\boldsymbol{\Gamma}_k'\boldsymbol{\Lambda}_k^{(a)}\mathbf{Z}_{k,i}')$. Integrating out all random effects simultaneously dramatically decreases the dimension of the integral needed for the marginal likelihoods. After specifying suitable priors for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_k$, we use the Laplace method to integrate over $(\boldsymbol{\beta}, \boldsymbol{\lambda}_k^{(a)}, \boldsymbol{\gamma}_k)$ to approximate the marginal likelihoods $p(\mathbf{Y}|M_k^{(a)})$ used to evaluate the Bayes factor $B_{10}^{(a)}$. For additional details regarding these multivariate t-distributions, see the Appendices.

As previously discussed, many of the existing methods for testing variance components are only applicable in simple settings, e.g. testing a single variance component. One major advantage of our approach is we can test multiple random effects simultaneously by modifying equation (11) such that the $\mathbf{Z}_{1,i}$, $\boldsymbol{\Lambda}_1^{(a)}$, $\boldsymbol{\Gamma}_1$, and $\mathbf{b}_{1,i}$ correspond to a model with several additional random effects. Incorporating default priors on the $\boldsymbol{\beta}$ coefficients, one can simultaneously compare models with varying numbers of both fixed and random effects.

3.3 Simulation study

We conduct a simulation study to test for the presence of a random slope. We define one predictor based on time, such that $\mathbf{x}_i = (1, 2, \dots, J)'$, $\mathbf{X}_i = (\mathbf{1}, \mathbf{x}_i)$, and $\boldsymbol{\beta} = (\beta_0, \beta_1)'$. Consistent with our previous notation, let M_1 refer to the random intercept model and M_2 refer to the random intercept and slope model. Letting $\mathbf{Z}_{1,i} = \mathbf{1}_J$, $\boldsymbol{\lambda}_1^{(a)} = \lambda_0^{(a)}$, and $\mathbf{b}_{1,i} = b_{i0}$, we have

$$M_1^{(a)} : y_{ij} = \beta_0 + \lambda_0^{(a)}b_{i0} + \beta_1x_{ij} + \varepsilon_{ij} \quad (12)$$

for the random intercepts model. Letting $\mathbf{Z}_{2,i} = (\mathbf{1}_J, \mathbf{x}_i)$, $\boldsymbol{\lambda}_2^{(a)} = (\lambda_0^{(a)}, \lambda_1^{(a)})'$, and $\mathbf{b}_{2,i} = (b_{i0}, b_{i1})'$, we have

$$M_2^{(a)} : y_{ij} = \beta_0 + (\lambda_0^{(a)} + \gamma_{12}x_{ij}\lambda_1^{(a)})b_{i0} + (\beta_1 + \lambda_1^{(a)}b_{i1})x_{ij} + \varepsilon_{ij} \quad (13)$$

for the random intercept and slope model. Our focus is to compare model $M_2^{(a)}$ to $M_1^{(a)}$. After integrating out $\mathbf{b}$ and σ^2 to produce marginal multivariate t-distributions, the integrals needed to calculate the marginal distributions $p(\mathbf{Y}|M_1^{(a)})$ and $p(\mathbf{Y}|M_2^{(a)})$ only have 3 or 5 dimensions, respectively. Hence the Laplace method can effectively be used to integrate over $(\beta_0, \beta_1, \lambda_0^{(a)})$ in $M_1^{(a)}$ and $(\beta_0, \beta_1, \lambda_0^{(a)}, \lambda_1^{(a)}, \gamma_{12})$ in $M_2^{(a)}$.

We simulated 100 data sets based on a random intercept and slope model under the standard notation of Laird and Ware (1982) as shown in (9), i.e.

$Y_{ij} = \beta_0 + b_{i0} + (\beta_1 + b_{i1})x_{ij} + \epsilon_{ij}$. We set $\beta_0 = 0$, $\beta_1 = 0.5$, $J = 10$, $\sigma^2 = 1$, and we generated the random effects from a multivariate normal distribution $\mathbf{b}_i \sim N_2(\mathbf{0}, \boldsymbol{\psi})$, in which $\psi_{11} = 1$ and $\psi_{12} = \rho(b_{i0}, b_{i1})\sqrt{\psi_{11}\psi_{22}} = -0.3$. We considered different combinations of the random slope variance component and sample size by varying $\sqrt{\psi_{22}} = 0, 0.04, 0.08, 0.15, 0.25$ across $n = 25, 50, 100, 500, 1000, 5000$. For implementing the Laplace approximation to the marginal likelihoods, we used prior distributions $\boldsymbol{\beta} \sim N(\mathbf{0}, \mathbf{I})$, $\sigma^2 \sim \text{InvGam}(1, 1)$, and $\gamma_{12} \sim N(0, 1)$.

[Table 2 about here.]

[Figure 2 about here.]

As illustrated in Table 2, our method performed well in favoring the correct model, but accuracy depended on both the sample size and the simulated variance of the random slope. In general, as the standard deviation of b_{i1} increased, our method increasingly favored $M_2^{(a)}$ over $M_1^{(a)}$. As the sample size increased, our method more accurately detected the absence of a random slope for $\sqrt{\psi_{22}} = 0$, and more accurately detected the presence of a random slope for $\sqrt{\psi_{22}} > 0$. For smaller sample sizes, our method generally detected the random slope for

$\sqrt{\psi_{22}} \geq 0.15$, indicating our method is useful even for small sample sizes with moderate to large random effects. Figure 2 shows box plots of $\log \hat{B}_{21}^{(1)}$ for $\sqrt{\psi_{22}} = 0$ and $\sqrt{\psi_{22}} = 0.08$. As n goes to infinity the estimated log Bayes factor $\hat{B}_{21}^{(1)}$ goes to infinity for $\sqrt{\psi_{22}} = 0.08$, and goes to negative infinity for $\sqrt{\psi_{22}} = 0$. This indicates the estimated Bayes factor increasingly favors the correct model as n increases.

As noted previously, the approximations to the marginal likelihoods do not seem to vary a great deal across parameterizations. Similar patterns were found in this simulation, with most differences extremely small. Occasionally we did observe large differences between the marginal likelihood estimates of $M_2^{(2)}$ and $M_2^{(1)}$ for simulated variances close to the boundary for $n = 5000$. It appears that this situation was due to occasional poor convergence of the maximization routine for the $\lambda_k^{(2)}$ parameterization, and not due to the Laplace approximation itself.

4. Illustrative examples

4.1 *Hamilton Rating Scale for Depression*

To illustrate our method, we consider a clinical trial of patients with bipolar I disorder (Calabrese et al., 2003), GlaxoSmithKline study SCAB2003. The investigators concluded that the treatment drug, lamotrigine, significantly delays the time to intervention for a depressive episode compared to placebo. The investigators also collected repeated measurements on the Hamilton Rating Scale for Depression (HAMD), a numerical measure of the severity of depressive symptoms. As a secondary analysis, we wish to determine if lamotrigine is effective in reducing depressive symptoms during the first year after randomization as measured by the HAMD-17 summary score. Larger HAMD-17 scores reflect higher levels of depression.

We consider 275 patients (160 lamotrigine 200/400 mg/day, 115 placebo) with at least one outcome measurement and complete covariate data. The number of repeated measurements

per subject ranges from 1 to 17, and HAMD-17 scores range from 0 to 35, with a mean value of 7. To better approximate normality, we used a square root transformation of HAMD-17 (sqrt-HAMD-17). We fit a linear mixed model with sqrt-HAMD-17 as the response, predicted by sqrt-HAMD-17 at screening and baseline, time (in years), treatment, gender, age (< 30, 30-40, 40-50, ≥ 50), and the number of depressive or mixed episodes in the last year (1-2 vs. ≥ 3). Screening refers to the time at enrollment, and baseline refers to the time of randomization (after stabilization).

In assessing the impact of lamotrigine on HAMD-17 scores, it is also interesting to assess the variability among patients with regards to the overall mean and slope. One might expect patients to have different patterns of depressive episodes across time, perhaps resulting from biological mechanisms or individual responses to drug treatment. This leads to the task of testing whether to include random effects in our model. Our focus is to compare models with varying combinations of a random intercept and slope, i.e. M_0 : a model without random effects; M_1 : a model with a random intercept; and M_2 : a model with a random intercept and slope. Based on the scale of both the response and the explanatory variables, we use vague priors on the fixed effects β and residual variance σ^2 that accommodate a wide range of reasonable mean values. We define these priors as $\beta \sim N_9(\mathbf{0}, 10\mathbf{I})$ and $\sigma^2 \sim \text{InvGamma}(0.01, 0.01)$.

The estimated log Bayes factors for the respective comparisons are $\log \hat{\beta}_{21} = 61.0$, $\log \hat{\beta}_{20} = 348.5$, $\log \hat{\beta}_{10} = 287.5$. These estimates show strong evidence for M_2 versus the other models, indicating the intercepts and slopes vary significantly by individual. We fit the preferred model, M_2 , using MCMC methods based on 15,000 samples, with a burn-in of 10,000. Figure 3 shows the predicted overall mean for each treatment group and the predicted individual sqrt-HAMD-17 for 50 random subjects. The predicted overall mean is based on a 40-50 year old female with 1-2 depressive episodes in the past year, and average values of sqrt-HAMD-

17 at screening and baseline (2.2 and 4.8, respectively). There is large variability in the subject specific intercepts and slopes. Lamotrigine use, age, and sqrt-HAMD-17 at screening and baseline all appear to be significant predictors of the outcome. A one unit increase in sqrt-HAMD-17 at baseline is associated with a 0.63 (95% CI = 0.51, 0.74) increase in mean sqrt-HAMD-17, and a one unit increase in sqrt-HAMD-17 at screening is associated with a 0.22 (95% CI = 0.02, 0.48) increase in mean sqrt-HAMD-17. On average, patients 30-40 years old have sqrt-HAMD-17 values 0.18 (95% CI = -0.20, 0.56) units greater than patients < 30 years old, patients 40-50 years old have sqrt-HAMD-17 values 0.47 (95% CI = 0.13, 0.82) units greater than patients < 30 years old, and patients ≥ 50 years old have sqrt-HAMD-17 values 0.39 (95% CI = 0.05, 0.73) units greater than patients < 30 years old. As the main association of interest, sqrt-HAMD-17 values for subjects on lamotrigine are on average 0.33 units lower (95% CI = -0.54, -0.10) than sqrt-HAMD-17 values for subjects on placebo. The 95% credible interval does not contain 0, indicating that lamotrigine may be effective at reducing depressive symptoms. These conclusions reinforce the time-to-event analysis of Calabrese et al. (2003).

[Figure 3 about here.]

4.2 *Exposure of disinfection by-products in drinking water and male fertility*

A multi-center study of 229 male patients from 3 sites (A, B, C) was conducted to evaluate the effect of disinfection by-products (DBP's) in drinking water on male reproductive outcomes in presumed fertile men. DBP exposure was measured using water system samples and data collected on individual water usage. Three exposure variables of interest for the outcome percent normal sperm are brominated haloacetic acids (HAA-Br), brominated trihalomethanes (THM-Br), and total organic halides (TOX).

Our focus is to model the response (% normal sperm) using the three DBP exposure variables. Because we are interested in each exposure's effect independent of the other

exposure variables, we fit three separate models, one for each DBP exposure. In each model we control for the following baseline covariates using indicator variables: male age (< 25 , $25-30$, $30-35$, > 35), education (high school or less, some college, graduated college), and the abstinence interval before taking the sample (2-3 days, 4-8 days, or > 8 days). We scale each predictor by subtracting the overall mean of the predictor and dividing by a constant c ($c = 10$ for HAA-Br and THM-Br, and $c = 100$ for TOX) to allow for better computational efficiency. We use a probit transformation of percent normal sperm and multiply the result by 5, so the transformed response has a range of -10.5 to -1.8, a mean of -5.6, and a variance of 1.8.

In assessing the impact of DBP's on sperm quality, it is also of interest to assess the variability among study sites with regards to the overall mean of percent normal sperm (i.e. intercept) and each DBP effect (i.e. slope). It may be the case that study site is a surrogate for unmeasured aspects of water quality or other unmeasured factors of interest. For each DBP exposure, we define three models based on the inclusion of random effects, i.e. M_0 : a model without random effects; M_1 : a model with a random intercept; and M_2 : a model with a random slope and intercept. Based on the scale of both the response and the explanatory variables, we use vague priors on the fixed effects β and residual variance σ^2 that accommodate a wide range of reasonable mean values. We define these priors as $\beta \sim N_9(\mu, \Sigma)$ and $\sigma^2 \sim \text{InvGamma}(0.01, 0.01)$, with $\mu = (-5.5, \mathbf{0}'_8)'$ and Σ a diagonal matrix with diagonal elements $(100, 10 \times \mathbf{1}'_8)$.

For HAA-Br, we observe moderate evidence for M_1 versus M_2 ($\hat{B}_{12} = 6.9$) and strong evidence for M_1 versus M_0 ($\log \hat{B}_{10} = 15.3$). For THM-Br, we observe strong evidence for M_1 versus both M_2 ($\hat{B}_{12} = 10.5$) and M_0 ($\log \hat{B}_{10} = 19.2$). For TOX, we observe weak evidence for M_1 versus both M_2 ($\hat{B}_{12} = 1.1$) and strong evidence for M_1 versus M_0 ($\log \hat{B}_{10} = 12.9$). Hence the random intercepts model M_1 is favored by the Bayes factor for all three DBP

exposures. For comparison, we fit both models M_1 and M_2 using MCMC methods based on 40,000 samples, with a burn-in of 40,000 for each model. We plot the predicted mean response based on M_2 , for a 30-35 year old male who has graduated college and has abstained for 2-3 days (Figure 4). For each of the three exposure models, one can see that there is some separation of the intercepts and varying degrees of agreement between the slopes. Although the point estimates of the slopes (based on the posterior means) appear to be quite different, the large variability associated with these estimates suggests that the slopes do not vary by study site. Hence we conclude that M_1 is the preferred model for each of the predictors. Based on M_1 , both HAA-Br and THM-Br have posterior distributions centered near 0, indicating little association between these DPB's and percent normal sperm. The posterior distribution of TOX tends to be centered below 0, with a posterior mean of -1.20 (95 % CI = -3.67,0.47); however, the 95% credible interval contains 0.

[Figure 4 about here.]

5. Discussion

We recommend our approach as a simple and efficient method in testing random effects in the linear mixed model. Our approach avoids issues with testing on the boundary of the parameter space, uses low-dimensional approximations to the Bayes factor, and incorporates default priors on the random effects. We have shown Laplace's method to be an effective approach to estimating Bayes factors, even in cases in which the variance of the random effect lies on the boundary. By scaling the random effects to the residual variance and introducing a parameter that controls the relative contribution of the random effects, we can effectively integrate out the random effects and reduce the dimensionality of the marginal likelihood. The scaling of the random effects to the residual variance makes the $\log N(\mathbf{0}, \mathbf{I})$ and $N(\mathbf{0}, \mathbf{I})$ distributions reasonable default priors for $\boldsymbol{\lambda}_k^{(1)}$ and $\boldsymbol{\lambda}_k^{(2)}$, respectively. Simulations

suggest that these priors have good small sample properties and consistency in large samples. Incorporating reasonable default priors on the fixed effects, our method can be used for comparing a large class of random effects models with varying fixed and random effects.

Alternative procedures for allowing default priors for model selection via Bayes factors are discussed by Berger and Pericchi (1996). These include the authors' proposed *intrinsic Bayes factors*, the Schwarz approximation (Schwarz, 1978), and the methods of Jeffreys (1961) and Smith and Spiegelhalter (1980). Gelman (2006) discuss various approaches to default priors specifically for variance components. Common approaches include the uniform prior (e.g. Gelman, 2007), the half- t family of prior distributions, and the inverse-gamma distribution (Spiegelhalter et al., 2003). These prior distributions can encounter difficulties when the variance components are close to 0. Other discussions of selecting default priors on variance components include Natarajan and Kass (2000), Browne and Draper (2006), and Kass and Natarajan (2006).

References

- Bartlett, M. S. (1957). Comment on "A Statistical Paradox" by D. V. Lindley. *Biometrika* **44**, 533–534.
- Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* **91**, 109–122.
- Browne, W. J. and Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models (with discussion). *Bayesian Analysis* **1**, 473–514.
- Cai, B. and Dunson, D. B. (2006). Bayesian covariance selection in generalized mixed models. *Biometrics* **62**, 446–457.
- Calabrese, J. R., Bowden, C. L., Sachs, G., Yatham, L. N., and Behnke, K. (2003). A placebo-controlled 18-month trial of lamotrigine and lithium maintenance treatment in

- recently depressed patients with bipolar I disorder. *Journal of Clinical Psychiatry* **64**, 1013–1024.
- Chen, Z. and Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics* **59**, 762–769.
- Commenges, D. and Jacqmin-Gadda, H. (1997). Generalized score test of homogeneity based on correlated random effects models. *Journal of the Royal Statistical Society, Series B* **59**, 157–171.
- Crainiceanu, C. M. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society, Series B* **66**, 165–185.
- Erkanli, A. (1994). Laplace approximations for posterior expectations when the mode occurs at the boundary of the parameter space. *Journal of the American Statistical Association* **89**, 250–258.
- Feng, Z. and McCulloch, C. E. (1992). Statistical inference using maximum likelihood estimation and the generalized likelihood ratio when the true parameter lies on the boundary of the parameter space. *Statistics & Probability Letters* **11**, 325–332.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis* **1**, 515–533.
- Gelman, A. (2007). Running WinBugs and OpenBugs from R. Available at www.stat.columbia.edu/~gelman/bugsR/.
- Genz, A. (1991). An adaptive numerical integration algorithm for simplices. In Sherwani, N. A., de Doncker, E., and Kapenga, J. A., editors, *Computing in the 90's, Proceedings of the First Great Lake Computer Science Conference*, pages 279–292, New York. Springer.
- Gilks, W. R. and Roberts, G. O. (1996). *Strategies for Improving MCMC*, page 96. Chapman and Hall, New York.

- Hsiao, C. K. (1997). Approximate Bayes factors when a mode occurs on the boundary. *Journal of the American Statistical Association* **92**, 656–663.
- Jeffreys, H. (1961). *Theory of Probability (3rd edition)*. Oxford University Press, Oxford, U.K.
- Kass, R. E. and Natarajan, R. (2006). A default conjugate prior for variance components in generalized linear mixed models (comment on article by Browne and Draper). *Bayesian Analysis* **1**, 535–542.
- Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.
- Kinney, S. K. and Dunson, D. B. (2007). Fixed and random effects selection in linear and logistic models. *Biometrics* (in press) .
- Kuonen, D. (2003). Numerical integration in S-plus or R: A survey. *Journal of Statistical Software* **8**, 1–14.
- Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.
- Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2005). Mixtures of g -priors for Bayesian variable selection. Technical Report ISDS 2005-12, Duke University, Department of Statistics and Decision Sciences, Durham, NC.
- Lin, X. (1997). Variance components testing in generalised linear models with random effects. *Biometrika* **84**, 309–326.
- Molenberghs, G. and Verbeke, G. (2007). Likelihood ratio, score, and Wald tests in a constrained parameter space. *The American Statistician* **61**, 22–27.
- Natarajan, R. and Kass, R. E. (2000). Reference Bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association* **95**, 227–237.
- Nelder, J. A. and Mead, R. (1965). A simplex algorithm for function minimization. *Computer*

Journal **7**, 308–313.

- Pauler, D. K., Wakefield, J. C., and Kass, R. E. (1999). Bayes factors and approximations for variance component models. *Journal of the American Statistical Association* **94**, 1242–1253.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.
- Self, S. G. and Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and the likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* **82**, 605–610.
- Shapiro, A. (1988). Towards a unified theory of inequality constrained testing in multivariate analysis. *International Statistical Review* **56**, 49–62.
- Silvapulle, M. J. (1992). Robust Wald-type tests of one-sided hypotheses in the linear model. *Journal of the American Statistical Association* **87**, 156–161.
- Sinharay, S. and Stern, H. S. (2001). Bayes factors for variance component models in generalized mixed models. In *Bayesian Methods with Applications to Science, Policy and Official Statistics*, pages 507–516. ISBA 2000 Proceedings.
- Smith, A. F. M. and Spiegelhalter, D. J. (1980). Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society, Series B* **42**, 213–220.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B* **64**, 583–640.
- Spiegelhalter, D. J., Thomas, A., Best, N. G., Gilks, W. R., and Lunn, D. (2003). WinBUGS User Manual, Version 1.4. Available at www.mrc-bsu.cam.ac.uk/bugs.
- Stram, D. O. and Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics* **50**, 1171–1177.

- Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistician* **81**, 82–86.
- Verbeke, G. and Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics* **59**, 254–262.
- Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology* **44**, 92–107.
- Zellner, A. (1986). *On assessing prior distributions and Bayesian regression analysis with g-prior distributions*, pages 233–243. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*.
- Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia (Spain)*, pages 585–603.
- Zhang, D. and Lin, X. (2007). *Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics*. *Model Uncertainty in Latent Variable and Random Effects Models (to appear)*.

Appendix A

The marginal distributions in Sections 2.1 and 3.2 can be derived from the integral

$$\begin{aligned}
 p(\mathbf{Y}|M_k^{(a)}) &= \int \int \left\{ \prod_{i=1}^n \int p(\mathbf{y}_i|\boldsymbol{\zeta}_k^{(a)}, b_i, \sigma^2) \pi(b_i) db_i \right\} \pi(\sigma^2) d\sigma^2 \pi(\boldsymbol{\zeta}_k^{(a)}) d\boldsymbol{\zeta}_k^{(a)} \quad (\text{A.14}) \\
 &= \int \int \left\{ \prod_{i=1}^n p(\mathbf{y}_i|\boldsymbol{\zeta}_k^{(a)}, \sigma^2) \right\} \pi(\sigma^2) d\sigma^2 \pi(\boldsymbol{\zeta}_k^{(a)}) d\boldsymbol{\zeta}_k^{(a)} \\
 &= \int \int p(\mathbf{Y}|\boldsymbol{\zeta}_k^{(a)}, \sigma^2) \pi(\sigma^2) d\sigma^2 \pi(\boldsymbol{\zeta}_k^{(a)}) d\boldsymbol{\zeta}_k^{(a)} \\
 &= \int p(\mathbf{Y}|\boldsymbol{\zeta}_k^{(a)}) \pi(\boldsymbol{\zeta}_k^{(a)}) d\boldsymbol{\zeta}_k^{(a)}.
 \end{aligned}$$

A multivariate t-distribution for a random vector $\mathbf{x}$ is typically denoted as $\mathbf{x}_{(p \times 1)} \sim t_p(d, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, in which

$$p(\mathbf{x}) = \Gamma\left(\frac{d+p}{2}\right) \frac{(\pi d)^{-p/2} |\boldsymbol{\Sigma}|^{-1/2}}{\Gamma(d/2)} \left\{ 1 + \frac{1}{d} (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}^{-\frac{d+p}{2}}. \quad (\text{A.15})$$

In (A.15), p is the dimension of $\mathbf{x}$, d is the number of degrees of freedom, $\boldsymbol{\mu}$ is the non-centrality parameter, and $\boldsymbol{\Sigma}$ is the covariance matrix. In order to express the marginal distributions $p(\mathbf{Y}|\boldsymbol{\zeta}_k^{(a)}, M_k^{(a)})$ in the form of (A.15), we must express the models in terms of the vector $\mathbf{Y} = (\mathbf{y}'_1 \dots, \mathbf{y}'_n)'$.

Appendix B

For the ANOVA model in Section 2.1, we can write $M_1^{(1)}$ in terms of $\mathbf{Y}$ as

$$M_1^{(1)} : \mathbf{Y} = \mu \mathbf{1}_m + \lambda \mathbf{W} \mathbf{b} + \boldsymbol{\varepsilon}, \quad (\text{B.1})$$

in which $\mathbf{1}_m$ is a $(m \times 1)$ vector of ones, $\mathbf{W}$ is a $(m \times n)$ block diagonal matrix of $(\mathbf{1}_{n_1}, \dots, \mathbf{1}_{n_n})$, $\mathbf{b} = (b_1, \dots, b_n)'$, and $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}'_1, \dots, \boldsymbol{\varepsilon}'_n)'$. It can then be shown that $(\mathbf{Y}|\mu, \lambda, M_1^{(1)}) \sim t_m(2v, \mu \mathbf{1}_m, \frac{w}{v} \boldsymbol{\Sigma}^{(1)})$, in which $\boldsymbol{\Sigma}^{(1)} = (\mathbf{I}_m + \lambda^2 \mathbf{W} \mathbf{W}')$. Similarly, M_0 can be expressed in terms of the vector $\mathbf{Y}$ as

$$M_0 : \mathbf{Y} = \mu \mathbf{1}_m + \boldsymbol{\varepsilon}. \quad (\text{B.2})$$

It can then be shown that $(\mathbf{Y}|\mu, M_0) \sim t_m(2v, \mu \mathbf{1}_m, \frac{w}{v} \mathbf{I}_m)$. It is also straightforward to show that $(\mathbf{Y}|\mu, \phi, M_1^{(2)}) \sim t_m(2v, \mu \mathbf{1}_m, \frac{w}{v} \boldsymbol{\Sigma}^{(2)})$, in which $\boldsymbol{\Sigma}^{(2)} = (\mathbf{I}_m + e^{2\phi} \mathbf{W} \mathbf{W}')$. For large datasets, the covariance matrix $\boldsymbol{\Sigma}$ may be too large to handle computationally in a mixed model setting. Hence it is preferable to express the multivariate t-distribution in terms of the subject-specific (independent) covariance matrices $\boldsymbol{\Sigma}_i$, as shown in equation (6)

Appendix C

For the linear mixed model in Section 3.2, $M_0^{(a)}$ can be expressed in terms of the vector $\mathbf{Y}$ as

$$M_0^{(a)} : \mathbf{Y} = \mathbf{X} \boldsymbol{\beta} + \mathbf{Z}_0 \mathbf{W}_0^{(a)} \mathbf{b}_0 + \boldsymbol{\varepsilon}, \quad (\text{C.1})$$

in which $\mathbf{X} = (\mathbf{X}'_1, \dots, \mathbf{X}'_n)'$ is a $(m \times p)$ design matrix, $\mathbf{Z}_0$ is a $(m \times q)$ block diagonal

matrix of $(\mathbf{Z}_{0,1}, \dots, \mathbf{Z}_{0,n})$, $\mathbf{b}_0 = (\mathbf{b}'_{0,1}, \dots, \mathbf{b}'_{0,n})'$ is a $(nq \times 1)$ vector of all random effects, $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}'_1, \dots, \boldsymbol{\varepsilon}'_n)'$, $\mathbf{W}_0^{(1)} = \mathbf{I}_n \otimes (\boldsymbol{\Lambda}_0^{(1)} \boldsymbol{\Gamma}_0)$ and $\mathbf{W}_0^{(2)} = \mathbf{I}_n \otimes (\boldsymbol{\Lambda}_0^{(2)} \boldsymbol{\Gamma}_0)$ are $(nq \times nq)$ matrices, in which $\otimes$ denotes the right Kronecker product (whereby the matrix on the right multiplies each element of the matrix on the left). We can express $M_1^{(a)}$ in terms of the vector $\mathbf{Y}$ as

$$M_1^{(a)} : \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1 \mathbf{W}_1^{(a)} \mathbf{b}_1 + \boldsymbol{\varepsilon}, \quad (\text{C.2})$$

in which $\mathbf{Z}_1$ is a block diagonal matrix of $(\mathbf{Z}_{1,1}, \dots, \mathbf{Z}_{1,n})$, $\mathbf{b}_1 = (\mathbf{b}'_{1,1}, \dots, \mathbf{b}'_{1,n})'$, $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}'_1, \dots, \boldsymbol{\varepsilon}'_n)'$, $\mathbf{W}_1^{(1)} = \mathbf{I}_n \otimes (\boldsymbol{\Lambda}_1^{(1)} \boldsymbol{\Gamma}_1)$, and $\mathbf{W}_1^{(2)} = \mathbf{I}_n \otimes (\boldsymbol{\Lambda}_1^{(2)} \boldsymbol{\Gamma}_1)$. It can then be shown that $p(\mathbf{Y}|\boldsymbol{\beta}, \boldsymbol{\lambda}_k^{(a)}, \gamma_k, M_k^{(a)}) \sim t_m(2v, \mathbf{X}\boldsymbol{\beta}, \frac{w}{v} \boldsymbol{\Sigma}_k^{(a)})$, in which $\boldsymbol{\Sigma}_k^{(a)} = (\mathbf{I}_m + \mathbf{Z}_k \mathbf{W}_k^{(a)} \mathbf{W}_k'^{(a)} \mathbf{Z}_k')$. As with testing a random intercept in the ANOVA setup, the covariance matrix $\boldsymbol{\Sigma}_k^{(a)}$ may be too large to handle computationally in a mixed model setting.

Figure 1: Box plot of $\log \hat{B}_{10}^{(1)}$, by ρ

Figure 2: Box plot of $\log \hat{B}_{21}^{(1)}$, by standard deviation of random slope

Figure 3: Predicted mean & individual HAMD-17, with random slope and intercept

* % normal sperm is on the scale of 5 times the probit
 ** Predictors are centered and scaled by a constant

Figure 4: Predicted mean of % normal sperm, with random slope and intercept

Table 1: Testing a random intercept, $\hat{B}_{10}^{(a)}$

n	ρ	Parameterization (1)						Parameterization (2)					
		Favor null			Favor random int.			Favor null			Favor random int.		
		< 0.1	0.1-0.33	0.33-1	1-3	3-10	> 10	< 0.1	0.1-0.33	0.33-1	1-3	3-10	> 10
25	0	6	67	22	4	0	1	7	71	18	3	0	1
	0.14	0	31	35	22	9	3	0	36	31	21	9	3
	0.33	0	1	12	20	25	42	0	1	14	18	25	42
	0.5	0	0	0	4	7	89	0	0	1	3	7	89
	0.69	0	0	0	0	0	100	0	0	0	0	0	100
50	0	23	50	18	7	2	0	27	51	13	7	2	0
	0.14	2	22	34	16	14	12	4	25	30	18	11	12
	0.33	0	0	7	4	8	81	0	0	7	5	8	80
	0.5	0	0	0	0	0	100	0	0	0	0	0	100
	0.69	0	0	0	0	0	100	0	0	0	0	0	100
100	0	43	43	12	2	0	0	52	36	10	2	0	0
	0.14	3	12	18	21	15	31	6	13	15	22	15	29
	0.33	0	0	1	0	3	96	0	0	1	0	4	95
	0.5	0	0	0	0	0	100	0	0	0	0	0	100
	0.69	0	0	0	0	0	100	0	0	0	0	0	100
500	0	75	18	7	0	0	0	82	14	4	0	0	0
	0.14	0	0	0	1	1	98	0	0	0	1	1	98
	0.33	0	0	0	0	0	100	0	0	0	0	0	100
	0.5	0	0	0	0	0	100	0	0	0	0	0	100
	0.69	0	0	0	0	0	100	0	0	0	0	0	100
1000	0	84	14	2	0	0	0	86	12	2	0	0	0
	0.14	0	0	0	0	0	100	0	0	0	0	0	100
	0.33	0	0	0	0	0	100	0	0	0	0	0	100
	0.5	0	0	0	0	0	100	0	0	0	0	0	100
	0.69	0	0	0	0	0	100	0	0	0	0	0	100
5000	0	96	3	1	0	0	0	96	3	1	0	0	0
	0.14	0	0	0	0	0	100	0	0	0	0	0	100
	0.33	0	0	0	0	0	100	0	0	0	0	0	100
	0.5	0	0	0	0	0	100	0	0	0	0	0	100
	0.69	0	0	0	0	0	100	0	0	0	0	0	100

* Table includes the percent of times that the estimated Bayes factors fell into the respective categories

Table 2: Testing for a random slope, $\hat{B}_{21}^{(a)}$

n	$\sqrt{\psi_{22}}$	Parameterization (1)						Parameterization (2)					
		Favor null			Favor random slope			Favor null			Favor random slope		
		< 0.1	0.1-0.33	0.33-1	1-3	3-10	> 10	< 0.1	0.1-0.33	0.33-1	1-3	3-10	> 10
25	0	97	2	1	0	0	0	97	2	1	0	0	0
	0.04	98	1	1	0	0	0	98	1	1	0	0	0
	0.08	85	6	5	3	1	0	85	6	5	3	1	0
	0.15	11	5	6	4	9	65	12	4	7	3	9	65
	0.25	0	0	0	0	1	99	0	0	0	0	1	99
50	0	100	0	0	0	0	0	100	0	0	0	0	0
	0.04	97	2	1	0	0	0	98	1	1	0	0	0
	0.08	65	11	8	7	1	8	65	11	8	7	1	8
	0.15	0	3	0	0	4	93	0	3	0	0	5	92
	0.25	0	0	0	0	0	100	0	0	0	0	0	100
100	0	99	1	0	0	0	0	100	0	0	0	0	0
	0.04	98	2	0	0	0	0	98	2	0	0	0	0
	0.08	37	11	14	10	9	19	37	11	15	9	10	18
	0.15	0	0	0	0	0	100	0	0	0	0	0	100
	0.25	0	0	0	0	0	100	0	0	0	0	0	100
500	0	100	0	0	0	0	0	100	0	0	0	0	0
	0.04	93	2	2	0	2	1	93	2	2	0	2	1
	0.08	0	0	1	1	0	98	0	0	1	1	0	98
	0.15	0	0	0	0	0	100	0	0	0	0	0	100
	0.25	0	0	0	0	0	100	0	0	0	0	0	100
1000	0	100	0	0	0	0	0	100	0	0	0	0	0
	0.04	77	11	8	2	1	1	78	11	7	2	1	1
	0.08	0	0	0	0	0	100	0	0	0	0	0	100
	0.15	0	0	0	0	0	100	0	0	0	0	0	100
	0.25	0	0	0	0	0	100	0	0	0	0	0	100
5000	0	100	0	0	0	0	0	100	0	0	0	0	0
	0.04	2	4	3	3	4	84	2	5	2	3	4	84
	0.08	0	0	0	0	0	100	0	0	0	0	0	100
	0.15	0	0	0	0	0	100	0	0	0	0	0	100
	0.25	0	0	0	0	0	100	0	0	0	0	0	100

* Table includes the percent of times that the estimated Bayes factors fell into the respective categories