The Schwarz Criterion and Related Methods for Normal Linear Models

BY DONNA K. PAULER

Department of Statistical Science, University College London,
Gower Street, London WC1E 6BT, U.K.

SUMMARY

The Schwarz criterion or Bayesian information criterion provides a simple consistent reference method for evaluating statistical models (Schwarz (1978)). In this paper we derive Schwarz’s criterion and two modifications for choosing fixed effects in Normal linear mixed models. The first modification allows an arbitrary, possibly informative prior for the parameter of interest. Replacing this prior with the Normal “unit-information” prior of Kass & Wasserman (1995) and the generalized Cauchy prior of Jeffreys (1961) yields the usual Schwarz criterion and a second modification, respectively. Under the null hypothesis, these criteria approximate Bayes factors using the corresponding priors to increased accuracy. In regression, the second modification also corresponds asymptotically to the Bayes factors of Zellner & Siow (1980), O’Hagan (1995), and is similar to the Bayes factor of Berger & Pericchi (1996). In mixed models, the effective sample size term in Schwarz’s formula is ambiguous due to correlation between observations. We propose an appropriate generalization of Schwarz’s approximation and apply our results to evaluate a large class of models for repeated neuron area measurements in alcoholic and suicidal patients.

Some key words: Bayes factor; Mixed effects models; Nested hypotheses; Regression; Schwarz criterion; Unit-information.

1 Introduction

For variable selection or more generally, for weighing alternative nested hypotheses in Normal linear models, many Bayesian reference methods have been derived and discussed. Recently developed methods exploit modern computational strategies such as Markov Chain Monte Carlo. In the context of regression, an abridged list of examples includes Zellner
& Siow (1980), Smith & Spiegelhalter (1980), Mitchell & Beauchamp (1988), George & McCulloch (1993), Gelfand & Dey (1994), Laud & Ibrahim (1995), Draper (1995), Kass & Wasserman (1995), O'Hagan (1995), and Berger & Pericchi (1996). Previous to these procedures, Schwarz proposed a very simple criterion, which we will denote by $S$, commonly referred to in the literature as the Bayesian information criterion or BIC (Schwarz (1978)). In this article we derive the appropriate form of $S$, propose two simple modifications, and justify their use for choosing Normal linear mixed models.

Schwarz's approximation has been widely discussed in the context of regression. See for example Smith & Spiegelhalter (1980), Nishii (1984), Haughton (1988), and Kass & Vaidyanathan (1992). Our contribution is to show that for tests of nested hypotheses, under the null hypothesis, $S$ is a specific case of a more general, higher-order approximation to the Bayes factor. This general approximation allows a possibly informative prior for the parameter of interest; $S$ corresponds to a particular choice of Normal “unit-information” prior (Kass & Wasserman (1995)). We show that several of the above Bayesian reference procedures are also special cases of this general approximation, corresponding to a generalization of the Cauchy prior recommended by Jeffreys (1961). $S$ and two modifications, allowing for an arbitrary or Cauchy prior, can be computed directly from standard regression software output. This along with their high accuracy make them highly competitive Bayesian reference procedures.

In linear mixed models the Schwarz criterion assumes a form similar to that for regression but the definition of the sample size $n$, which is the order of the Fisher information, is ambiguous due to correlation between observations. For example, in a growth curve analysis where observations consist of measurements obtained at $T$ time points from a group of $I$ individuals, a test for the between-subject effect of sex depends only on the number of individuals, $I$, and not the total number of observations, $IT$.

The remainder of this paper proceeds as follows. In Section 2 we define the Bayes factor, posterior probability and Schwarz approximation for comparing nested models and classes of models. In Section 3 we derive a general form of Schwarz’s approximation and show its relation to the usual Schwarz criterion, $S$, and some of the exact Bayesian procedures mentioned above. In Section 4 we extend results to mixed models. In the reference case, we
apply our results to evaluate a large class of models for repeated neuron area measurements in a group of alcoholic and suicidal patients.

2 Definitions

The Schwarz criterion is an asymptotic approximation to the Bayes factor, a consistent procedure for model determination under suitably regular prior distributions and likelihoods (Doob (1949)). In this section, we define the Bayes factor and Schwarz criterion for comparison of nested hypotheses.

Consider an observation vector \( y \) of dimension \( n \) coming from a distribution \( p(y|\psi, \eta) \) where the parameters \( (\psi, \eta) \) satisfy \( \psi \in \mathbb{R}^{p_0} \) and \( \eta \in \mathbb{R}^{p_1} \). Suppose interest lies in the test of the model, \( M_0 : \psi = \psi_0 \) versus \( M_1 : \psi \in \mathbb{R}^{p_0} \). Denote by \( \text{pr}(M_0) \) and \( \text{pr}(M_1) \) the prior probabilities of models \( M_0 \) and \( M_1 \), respectively. The Bayes factor is the ratio of posterior to prior odds of \( M_0 \) and provides evidence in favor of \( M_0 \) expressed by the data, \( y \). It can be written as

\[
BF = \frac{\text{pr}(M_0|y)/\text{pr}(M_1|y)}{\text{pr}(M_0)/\text{pr}(M_1)} = \frac{\int p(y|\psi_0, \eta)\pi_0(\eta)d\eta}{\int p(y|\psi, \eta)\pi(\psi, \eta)d\psi d\eta}
\]

where \( \pi_0(\eta) \) and \( \pi(\psi, \eta) \) are the respective priors for parameters under \( M_0 \) and \( M_1 \). For expressing evidence in favor of or against a model, the Bayes factor is desirable because of its direct interpretation (Jeffreys (1961)). However, the Bayes factor requires specification of proper priors that may be perceived as subjective or ad hoc. We discuss this issue in Section 3. In addition, as seen by (2.1), the Bayes factor requires evaluation of possibly high-dimensional integrals. Recent sampling approaches to calculation of these integrals can be found in Verdinelli & Wasserman (1995), DiCiccio et al. (1995), Chib (1995), Raftery (1996b), and two recent reports by Geman and Meng (1996) and Meng and Wong (1996).

The Schwarz criterion provides a very simple approximation to the Bayes factor given by

\[
S = \ell_0(\hat{\eta}_0) - \ell(\hat{\psi}, \hat{\eta}) + \frac{p_0}{2} \log n
\]

where \( \hat{\eta}_0 \) maximizes the null-hypothetical loglikelihood \( \ell_0(\eta) = \log p(y|\psi_0, \eta) \) and \( (\hat{\psi}, \hat{\eta}) \) maximizes the unrestricted loglikelihood \( \ell(\psi, \eta) \). As seen by (2.2), \( S \) is a simple function...
of the familiar likelihood ratio test statistic and the priors $\pi_0$ and $\pi$ do not appear in its formula.

Raftery (1996a) used Schwarz’s criterion to approximate posterior probabilities for evaluating large classes of generalized linear models. Suppose there are a collection of $K$ models, $M_1, \ldots, M_K$, with $M_1$ the smallest model nested in all of the others. Then relation (2.1) shows that the posterior probability of $M_k$ is given by:

$$\text{pr}(M_k | Y) = \frac{\alpha_k BF_{k1}}{\sum_{k=1}^{K} \alpha_k BF_{k1}}$$

(2.3)

where $\alpha_k = \text{pr}(M_k)/\text{pr}(M_1)$ and $BF_{k1}$ is the Bayes factor in favor of $M_k$ over $M_1$ for $k = 1, \ldots, K$. Schwarz’s criterion can be used to approximate the Bayes factors appearing in (2.3).

3 Regression

3.1 Arbitrary Priors

The Schwarz criterion is applicable for testing coefficients in Normal linear regression. In this section we derive a generalization of (2.2) which incorporates an arbitrary, possibly informative prior, and corresponds to the Bayes factor (2.1) to increased accuracy, $O_P(n^{-1/2})$. We later show how the generalization reduces to (2.2) for a specific intuitively reasonable choice of default prior.

To define the nested regression problem, let $Y = (Y_1, \ldots, Y_n)$ where $Y_i = x_i'\gamma + e_i$, $x_i$ is a given vector in $\mathbb{R}^p$, $e_i \sim N(0, \sigma^2)$ independently for $i = 1, \ldots, n$, $\sigma \in \mathbb{R}^+$, and $\gamma \in \mathbb{R}^p$. Denote the Normal probability density function for $Y$ by $p(\cdot | \gamma, \sigma)$, suppressing the known matrix of covariates $X_n = [x_1, \ldots, x_n]$. Setting $\gamma' = (\psi', \beta')$ with dim $\psi = p_0$, we are interested in ascertaining whether $M_0 : \psi = \psi_0$ is true versus the alternative $M_1 : \psi \in \mathbb{R}^{p_0}$. The Bayes factor for this problem is given by (2.1) with the vector of nuisance parameters, $\eta = (\beta, \sigma)$.

Schwarz’s approximation to the Bayes factor is derived by applying Laplace’s method to the numerator and denominator of (2.1). The version of Laplace’s method we use replaces the probability density $p(\cdot | \psi, \beta, \sigma)$ in these integrals with a Normal density centered at the maximum likelihood estimate with variance equal to the inverse observed information, $\hat{V} = (-D^2 \ell(\hat{\psi}, \hat{\beta}, \hat{\sigma}))^{-1}$. Here, $D^2 \ell$ denotes the second derivative of the loglikelihood. See
Chen (1985) and Kass et al. (1990) for a full discussion of the Laplace method. Kass et al. (1990) show that to apply the Laplace approximation and have it hold with probability one it is sufficient for the family of probability densities, $p(\psi, \beta, \sigma)$, to be Laplace regular. In regression the following assumption must hold.

**ASSUMPTION 1** For $n$ sufficiently large, if $\lambda_{n1}$ and $\lambda_{np}$ are the largest and smallest eigenvalues of $n^{-1}X_n^T X_n$ then there exists $a, b > 0$ such that $a < \lambda_{np}, \lambda_{n1} < b$.

We assume the following four structural and regularity conditions.

**CONDITION 1** Assume the Fisher information matrix, $I_{\psi,\eta}$, is null-orthogonal between $\psi$ and $\eta$ (Kass & Vaidyanathan (1992)). In other words, $I_{\psi,\eta}$ is orthogonal or block diagonal under $M_0$.

**CONDITION 2** Assume the marginal prior for $\eta$ is the same under both hypotheses. That is, denoting the prior under $M_0$ by $\pi_0(\eta)$, the prior under $M_1$ has the form $\pi(\psi, \eta) = \pi_0(\eta)\pi_{0\psi}(\psi|\eta)$.

**CONDITION 3** Let $B_n$ denote the Fisher information under the true model from a sample of size $n$ and $I_n^*$ the observed information. Let $\epsilon > 0$. Then

$$\lim_{n \to \infty} P \left( \sup_{||\gamma||=1} \left| 1 - \gamma' B_n^{-1/2} I_n^* B_n^{-1/2} \gamma \right| < \epsilon \right) = 1.$$

**CONDITION 4** Finally, suppose the maximum likelihood estimate $\hat{\psi}$ under the alternative satisfies $\hat{\psi} - \psi_0 = O_p(n^{-1/2})$ as it would if the true value of $\psi$ were either $\psi_0$ or a neighboring alternative $\psi_n$ such that $\psi_n - \psi_0 = O(n^{-1/2})$.

We believe Conditions 1 and 2 entail no substantial loss of generality for regression models. For Condition 1, let $X_{n0}$ denote the $n \times p_0$ submatrix of $X_n$ corresponding to $\psi$ and $X_{n1}$ the $n \times (p-p_0)$ submatrix corresponding to $\beta$. It is always possible to orthogonalize the design matrix by regressing the columns of $X_{n0}$ on $X_{n1}$ and taking the residuals to be the new $X_{n0}$. Interpretation of $\psi$ remains the same so prior specification for $\psi$ need not be altered. Condition 2 is justified based on Condition 1. Condition 3 requires that the observed information converge to the Fisher information in probability. In regression, this
condition follows directly from Assumption 1. Finally, when Condition 4 does not hold, the Bayes factor is exponentially small with probability one for large samples and it no longer makes much practical difference whether an approximation is accurate or crude. Let \(a_n\) be the point where the Schwarz approximation to the Bayes factor, \(BF\), breaks down and \(b_n\) a point past which \(BF\) is so small that an accurate approximation is not needed. A referee has pointed out that Condition 4 is reasonable only if the prior probability of observing data in the interval \([a_n, b_n]\) approaches zero as \(n\) increases.

The above conditions lead to the first modification of \(S\):

\[
S_A = S - p_0/2 \log 2\pi + 1/2 \log |I_\psi(\psi_0, \hat{\beta}, \hat{\sigma})| - \log \pi_{\psi|\beta,\sigma}(\psi|\hat{\beta}, \hat{\sigma})
\]

where \(I_\psi(\psi, \beta, \sigma) = X'_\psi X_\psi/(n\sigma^2)\) denotes the Fisher information in a sample of size one for \(\psi\), and \(|\cdot|\) denotes determinant. Given an informative prior, \(\pi_{\psi|\beta,\sigma}(\psi|\beta, \sigma)\) for \(\psi\), \(S_A\) is computable directly from output of most standard regression packages. Note also that \(S_A\) does not depend on the prior for the nuisance parameters, \((\beta, \sigma)\). The next proposition shows that \(S_A\) is a high-order approximation to a Bayes factor, \(BF_A\), using prior \(\pi_{\psi|\beta,\sigma}(\psi|\beta, \sigma)\) for \(\psi\) and any suitably regular prior for \((\beta, \sigma)\).

**Proposition 3.1** Suppose Assumption 1 holds, \(\pi_0\) and \(\pi\) are four times continuously differentiable and chosen so that the integrals in (2.1) are finite, and Conditions 1 - 4 hold. Then, \(\log BF_A - S_A = O_P(n^{-1/2})\).

**Proof:** See Appendix. \(\square\)

### 3.2 Default Priors

In practice there may be little or no a priori information for \(\psi\). For example, when there are large numbers of potential covariates, specification of priors for the many choices of \(\psi\) becomes infeasible. For these cases it is desirable to use “noninformative” priors which depend only on structural assumptions of the model. There are some minimal requirements and desiderata for use of these priors in Bayes factors (Jeffreys (1961)). The first requirement, and perhaps the most restrictive, is that the prior for \(\psi\) must be proper. Under improper priors, the marginal densities appearing in (2.1) may be infinite or proportional to any arbitrary constant \(c\) depending on the improper prior specification, \(\pi(\psi) \propto cd\psi\).
Second, to avoid unfairness towards the alternative, the prior for $\psi$ under $M_1$ should be located near the null value $\psi = \psi_0$. Finally, as in any reference prior formulation, the prior variance for $\psi$ should represent vague in formation that is calibrated with, but clearly less than the information for $\psi$ in the likelihood. One choice of prior variance is the inverse Fisher information for $\psi$ in the likelihood arising from a sample of size one. This was the intuition behind the “unit-information” prior studied by Kass & Wasserman (1995) and discussed previously by Smith & Spiegelhalter (1980). Use of a conjugate Normal density leads to the following “unit-information” prior for $\psi$:

$$\pi_{\psi|\beta,\sigma}(\psi|\beta,\sigma) = N_p(\psi_0, I^{-1}_{\psi\psi}(\psi_0, \beta, \sigma)), \quad (3.2)$$

where $N_p(a, B)$ is the $p$-variate Normal density with mean $a$ and variance-covariance matrix $B$. The following theorem shows that $S$ corresponds to the Bayes factor using (3.2) to increased accuracy, $O_P(n^{-1/2})$.

**THEOREM 3.1** Assume the conditions of Proposition 3.1 hold. Denote by $BF_N$ the Bayes factor using prior (3.2) for $\psi$ under $M_1$. Then $S - \log BF_N = O_P(n^{-1/2})$.

**PROOF:** The multivariate Normal prior for $\psi$ evaluated at the maximum likelihood estimate is given by

$$\pi_{\psi|\beta,\sigma}(\hat{\psi}, \hat{\beta}, \hat{\sigma}) = (2\pi)^{-p_0/2}|I_{\psi\psi}(\psi_0, \hat{\beta}, \hat{\sigma})|^{1/2}\exp\left\{-\frac{1}{2}(\hat{\psi} - \psi_0)'I^{-1}_{\psi\psi}(\hat{\psi} - \psi_0)\right\}$$

$$= (2\pi)^{-p_0/2}|I_{\psi\psi}(\psi_0, \hat{\beta}, \hat{\sigma})|^{1/2}\left\{1 + O_P(n^{-1})\right\},$$

where the last equality follows by a Taylor series expansion to the final exponent term, $\hat{\psi} - \psi_0 = O_P(n^{-1/2})$ and $I_{\psi\psi}(\psi_0, \hat{\beta}, \hat{\sigma}) = O_P(1)$. Substituting the inverse of this expression into (3.1) yields the result. $\square$

Jeffreys (1961) recommended the same location and scale in the prior for $\psi$ when testing $M_0 : \psi = \psi_0$ in the one-dimensional case but used the $t$-distribution with one degree of freedom, or Cauchy density, instead of the Normal density. Jeffreys preferred the Cauchy density as the simplest density satisfying specific desiderata of the Bayes factor, some of which are different than those discussed here (Jeffreys (1961), pgs. 268-273). Jeffreys’ prior may be extended for tests of multiple parameters by defining a multivariate $t$-distribution with one degree of freedom with the same location and scale as above:

7
\[ \pi_{\psi|\beta, \sigma}^c(\psi|\beta, \sigma) = \text{Cauchy}(\psi_0, I^{-1}_{\psi\psi}(\psi_0, \beta, \sigma)). \] (3.3)

Substituting prior (3.3) into (3.1) yields the following simple modification of $S$.

**Corollary 3.1** Under the conditions of Theorem 3.1, replace the prior (3.2) by (3.3). Define $S_C$ by

\[ S_C = S + \log r \] (3.4)

where $r = [\pi^{1/2}/\{2^{m_1/2}\Gamma(p_0/2 + 1/2)\}]$ and $BF_C$ the Bayes factor using prior (3.3). Then $S_C - \log BF_C = O_P(n^{-1/2})$.

**Proof:** Similar to Theorem 3.1, replacing $\pi$ by $\pi^c$. \square

Note that the last term in (3.4) is monotonically decreasing in $p_0$ so that $S_C$ favors the alternative more than $S$ as the difference in dimensionality between the null and alternative models increases.

### 3.3 Other Reference Methods

Other forms of reference priors for $\psi$ leading to alternative Bayes factors and approximations have recently been proposed in the literature. As we show here, under the null hypothesis many of these methods correspond asymptotically to $S_C$.

Zellner & Siow (1980) used the same Cauchy prior $\pi^c$ but applied Laplace's approximation only to the integral for $\psi$. Smith & Spiegelhalter (1980) generalized the prior variance in the Normal prior (3.2) to be $\sigma^2\rho(n)(X^T_0X_0)^{-1}$ so that $\rho(n)$ determines the amount of shrinkage. Then they applied Laplace's approximation to obtain (2.2).

Recently, Berger & Pericchi (1996) and O'Hagan (1995) proposed adaptations of Bayes factors to cope with improper priors. Berger & Pericchi (1996) placed an improper reference prior on $\psi$ and used the smallest subsample of the data that makes the prior proper to update the prior. To adjust for arbitrary choices of the minimal sample their arithmetic intrinsic Bayes factor (AIBF) averages over all possible minimal samples. O'Hagan (1995) defined the fractional Bayes factor (FBF) by dividing the Bayes factor (2.1) by the corresponding
Bayes factor using only a fraction of the likelihood in the numerator and denominator: $p(y|\psi_0, \eta)^b$ and $p(y|\psi, \eta)^b$, respectively. This causes the arbitrary constant in the improper prior specification for $\psi$ to cancel. O’Hagan suggested $b$ equal to $m/n$ where $m$ is the minimal sample size defined by Berger & Pericchi (1996) and $n$ is the total sample size but gave alternatives when robustness is a concern.

For testing $M_0 : \psi = \psi_0$ versus $M_1 : \psi \in \mathbb{R}^{p_0}$ in the regression model of Section 3.1, the following result shows that under $M_0$ the reference methods of Zellner & Siow (1980), which we will denote by ZS, O’Hagan (1995), and $S_C$ essentially correspond to Bayes factors using the same Cauchy prior, $\pi^c$, given in Corollary 3.1. Berger & Pericchi (1996) show that the AIBF uses a similar Cauchy intrinsic prior. Under $M_1$ FBF and ZS also correspond to real Bayes factors which agree asymptotically.

**RESULT 3.1** For FBF use the reference priors $\pi_0(\beta, \sigma) \propto 1/\sigma d\beta d\sigma$ under $M_0$, $\pi(\gamma, \sigma) \propto 1/\sigma d\gamma d\sigma$ under $M_1$ and fraction $b = (p + 1)/n$. Then $\text{FBF} - \text{ZS} = O_P(n^{-1})$ and if further, $M_0$ is true, $S_C - \text{ZS} = O_P(n^{-1})$ where $S_C$, ZS, and FBF are the log Bayes factors in favor of the null hypothesis, $M_0 : \psi = \psi_0$.

**PROOF:** See Appendix. $\square$

### 3.4 One-Way Analysis of Variance

Box & Tiao (1973) pgs. 246-247 present two data sets consisting of observations from 6 groups with 5 replications in each group. The first example is yield of dyestuff in grams of standard color coming from 6 batches. The between-batches mean square is large compared to the within-batches mean square suggesting that the null hypothesis of no difference between the batches is false, or that Condition 4 fails. The second example is artificially constructed so that Condition 4 holds. Treatng batch effects as fixed, the model is given by $y_{ij} = \beta + \psi_i + \epsilon_{ij}$ where $\epsilon_{ij} \sim N(0, \sigma^2)$ are independent for $i = 1, ..., 6$ and $j = 1, ..., 5$, and $\psi_0 = -\Sigma_{i=1}^5 \psi_i$. A test of the null hypothesis of no batch effect corresponds to testing $M_0 : \psi = 0$ versus $M_1 : \psi \in \mathbb{R}^5$ where $\psi = (\psi_1, ..., \psi_5)$. The Normal unit-information prior (3.2) for $\psi$ is given by $\pi_{\psi|\sigma}(\psi|\sigma) = N_5(0_5, 6\sigma^2[I_5 + J_5])$ where $0_5$ is a vector of zeroes of length 5, $I_5$ is the $5 \times 5$ identity matrix, and $J_5$ is a $5 \times 5$ matrix of ones.

To compare the various reference methods described in Section 3.3 we show results in
Table 1 listed as Bayes factors in favor of the alternative. For instance, in Example 1, the value of $S$ indicates the posterior odds of the alternative model were nearly 5 times as high as the prior odds. We include $S$, $S_C$, FBF using the fraction $b = 7/30$, ZS, and AIBF. In calculations of Bayes factors FBF and AIBF, the reference prior $\pi(\beta, \sigma) \propto 1/\sigma d\beta d\sigma$ is used under $M_0$ and $\pi(\gamma, \sigma) \propto 1/\sigma d\gamma d\sigma$ under $M_1$. We also show the exact Bayes factor, $BF_N$, using the Normal unit-information prior and marginal prior $\pi_0(\beta, \sigma) \propto 1/\sigma d\beta d\sigma$ under $M_0$ and $M_1$.

Ex. 1 Dyestuff  Ex. 2 Simulated

<table>
<thead>
<tr>
<th></th>
<th>Ex. 1</th>
<th>Ex. 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$-value</td>
<td>0.004</td>
<td>0.731</td>
</tr>
<tr>
<td>$S$</td>
<td>4.836</td>
<td>0.001</td>
</tr>
<tr>
<td>$S_C$</td>
<td>30.855</td>
<td>0.007</td>
</tr>
<tr>
<td>ZS</td>
<td>5.129</td>
<td>0.008</td>
</tr>
<tr>
<td>FBF</td>
<td>4.424</td>
<td>0.007</td>
</tr>
<tr>
<td>AIBF</td>
<td>5.360</td>
<td>0.015</td>
</tr>
<tr>
<td>$BF_N$</td>
<td>2.047</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Table 1: Bayes Factor in Favor of the Alternative. $S$: Schwarz criterion, $S_C$: Schwarz criterion with Cauchy prior, ZS: Zellner and Siow’s Bayes factor, FBF: O’Hagan’s fractional Bayes factor, AIBF: arithmetic intrinsic Bayes factor, $BF_N$: Bayes factor using Normal unit-information prior.

Example 2 was constructed so that $M_0$ is most likely true and the assumptions of Theorem 3.1 and Result 3.1 hold. Therefore, we consider this example first. Accordingly, $S$ is practically identical to $BF_N$ despite the small sample size and is approximately an order of magnitude lower than $S_C$. By (3-4) $S_C$ is 6.4 times as high as $S$ for these examples. $S_C$, ZS, and FBF all agree quite closely as predicted by Result 3.1. The AIBF has a value similar to $S_C$. In regression the Bayes factors discussed here can be expressed as functions of both the $p$-value of the observed $F$-statistic for testing $\psi = 0$ and the sample size. To see
how quickly these Bayes factors agree with sample size, we plotted them on the log base 10 scale for the $p$-value in this example in Figure 1b. ZS and FBF agree almost immediately at the minimal sample size of seven. $S_C$ converges to ZS and $S$ converges to $BF_N$ rather quickly.

![Figure 1a: P-value = .0044](image)

![Figure 1b: P-value = .7311](image)

Figure 1: Box & Tiao examples: Bayes factors in favor of alternative on the log base 10 scale for two $p$-values of the observed $F$-statistic. $S$: Schwarz criterion, $S_C$: Schwarz criterion with Cauchy prior, ZS: Zellner and Siow’s Bayes factor, FBF: O’Hagan’s fractional Bayes factor, and $BF_N$: Bayes factor using Normal unit-information prior.

In Example 1 $S_C$ and $S$ overestimate evidence in favor of $M_1$ compared to ZS and $BF_N$, respectively. As predicted by Result 3.1, ZS and FBF are close. Figure 1a shows a plot analogous to Figure 1b for the $p$-value observed in Example 1. For small sample sizes,
S and $S_C$ grossly overestimate their respective Bayes factors, whereas ZS and FBF agree almost immediately.

4 Mixed Models

4.1 Sample Size

Schwarz’s criterion may also be applied to tests of fixed effects in mixed linear models with similar arguments as in regression. However, in the default case its formula (2.2) requires a more precise definition. The linear mixed model assumes the following form (Laird & Ware (1982)). For the $i$th unit, $i = 1, ..., I$,

$$Y_i = X_i \gamma + Z_i b_i + \varepsilon_i \quad (4.1)$$

where $Y_i$ is a $n_i \times 1$ vector of observations, $X_i$ and $Z_i$ are $n_i \times p$ and $n_i \times q$ design matrices, respectively, $\gamma$ is a $p \times 1$ vector of fixed effects, $b_i$ is a $q \times 1$ vector of random effects such that $b_i \sim N(0, D)$, and $\varepsilon_i$ is a $n_i \times 1$ vector such that $\varepsilon_i \sim N(0, \sigma^2 I_{n_i})$. The $b_i$ and $\varepsilon_i$ are each independent and identically distributed and independent of each other for $i = 1, ..., I$. Note, we have assumed that within-subject errors, $\varepsilon_{ij}$, are independent and identically distributed $N(0, \sigma^2)$. A more general formulation allows $\varepsilon_i \sim N(0, R_i)$, where $R_i$ may be specified as an autoregressive, spatial or unstructured variance-covariance matrix. We discuss these types of models later in this section. Marginally the $Y_i$ are independent Normal random variables with mean $X_i \gamma$ and variance $V_i = V_i(\delta) = Z_i D Z_i^T + \sigma^2 I_{n_i}$ for $i = 1, ..., I$. We denote the unknown variance components of $V_i$, assumed to be constant across observations, by $\delta$ and the total number of observations by $N = \sum_{i=1}^I n_i$.

To understand the ambiguity in direct extension of Schwarz’s formula (2.2) for mixed models, it helps to review how it was defined for regression. The last term appearing in (2.2), which we will refer to as the penalty term, is derived from the order of the determinant of Fisher information for the parameter of interest, $\psi$. In regression, elements of the information matrix for $\psi = (\psi_1, ..., \psi_{p_0})$ are given by $-\frac{\partial^2 I}{\partial \psi_k \partial \psi_m} = \frac{x_k^T x_m}{\sigma^2}$, for $k, m = 1, ..., p_0$, where $x_k$ is the column of the design matrix, $X_{n0}$, corresponding to $\psi_k$. Assumption 1 of Section 3.1 requires that $x_k^T x_k = O(n)$ for $k = 1, ..., p_0$ and hence the determinant of information is of order $n^{p_0}$. Note that in the examples of Section 3.4 Assumption 1 requires
hypothetical sampling from all $I$ groups, with proportion of total sample size bounded from zero.

In model (4-1) elements of the information matrix for $\psi$ are given by

$$-rac{\partial^2 \ell}{\partial \psi_k \partial \psi_m} = \sum_{i=1}^{l} x_{ik} V_i^{-1} x_{im} \quad k, m = 1, \ldots, p_0,$$

where $X_i = [x_{i1} \cdots x_{ip_0}]$ and $V_i^{-1} = \frac{1}{\sigma^2} \left[ I_{n_i} - \frac{1}{\sigma^2} Z_i D (I_q + \frac{1}{\sigma^2} Z_i^T D^{-1} Z_i) \right]$ for $i = 1, \ldots, I$ (Harville (1977)). The order of these elements depend intimately on $V_i$ and vary for different elements, $\psi_k$. To see this assume that for each $i$, the columns of $X_i$ are orthogonal and that $Z_i$ is composed of a subset of the columns of $X_i$: $Z_i = [x_{i1} \cdots x_{is}]$ for $s \leq p_0$ and $i = 1, \ldots, I$. That is, the first $s$ effects follow a Normal distribution and the remaining $p_0 - s$ are fixed. Assume also that the variance-covariance matrix for random effects is diagonal; $D = \text{diag}(\sigma_1^2, \ldots, \sigma_q^2)$. Then formulas for the diagonal elements of (4-2) can be written explicitly as $-rac{\partial^2 \ell}{\partial \psi_k^2} = \sum_{i=1}^{l} \frac{x_{ik}^2}{\sigma^2}$ for $k = 1, \ldots, s$ and $-rac{\partial^2 \ell}{\partial \psi_k^2} = \sum_{i=1}^{l} \frac{x_{ik}^2}{\sigma^2}$ for $k = s + 1, \ldots, p$. Assuming that $x_{ik}^2 \sim O(n_i)$ for $k = 1, \ldots, p_0, i = 1, \ldots, I$, the information is of order $I$, the number of subjects, for fixed effects with associated random effects, and of order $N$, the total number of observations, for fixed effects with no associated random effects. This shows that the penalty term appearing in (2-2) will depend on which parameters are tested and the specific variance-covariance structure of the model.

Similar arguments as above hold for nonorthogonal $X_i$ and $Z_i$, and nondiagonal $D$ but expressions become unwieldy. Lemma 5.1 in the Appendix shows an analytic derivation of orders for these cases. Note, we have assumed that random effects are a subset of fixed effects and that within-subject errors are independent. For other cases, such as when random effects are a linear combination of fixed effects or the variance-covariance matrix of within-subject errors is nondiagonal, the order of Fisher information is not easily decomposed into effects which do and do not have reduced information. In these cases the specific model needs to be analyzed for sample size determination or a more general approximation depending only on an arbitrary prior implemented. We derive this general approximation in Section 4-2.
4.2 Schwarzs’s Modification

In model (4.1) assume $\gamma = (\psi, \beta)$ and consider again the test of the nested hypothesis, $M_0 : \psi = \psi_0$ versus $M_1 : \psi \in \mathbb{R}^{p_0}$. Denote the prior under $M_0$ by $\pi_0(\beta, \delta)$ and the prior under $M_1$ by $\pi(\gamma, \delta)$. The Bayes factor for this test is given by (2.1), replacing $\eta$ by $(\beta, \delta)$.

Laplace’s method can be applied to the numerator and denominator integrals of the Bayes factor with error depending on the smallest eigenvalue of the Fisher information matrix, $I_{\gamma, \delta}^N$. See Chen (1985) or the unpublished thesis, Carnegie Mellon University (Pauer (1996)). Regularity conditions for Laplace’s approximation are those required for joint asymptotic Normality in mixed models (Miller (1977), Mardia & Marshall (1984), Cressie & Lahiri (1993)). For the remainder of this paper we will assume these conditions hold and collect them into the following assumption.

**Assumption 2** The regularity conditions required for the priors $\pi_0$, $\pi$ and asymptotic Normality conditions for the log likelihood for model (4.1) hold (Chen (1985)).

Next assume Conditions 1 – 4 of Section 3.1 hold. In Condition 4, $n$ now refers to the smallest eigenvalue of the Fisher information matrix. For Condition 1, note that the information is always orthogonal between $\gamma$ and $\delta$. Also, $\beta$ can be made null orthogonal to $\psi$ by the transformation $\beta \rightarrow \alpha = \beta + (X_\beta' V^{-1} X_\beta)^{-1} X_\beta' V^{-1} X_\psi \psi$ where $X_\beta$ and $X_\psi$ are portions of the $N \times p$ design matrix $X$, which concatenates the $n_i \times p$ design matrices $X_i$ for $i = 1, \ldots, I$, corresponding to $\beta$ and $\psi$, respectively, and $V = \text{diag}(V_1, \ldots, V_I)$. These assumptions lead to the modified Schwarz approximation given by:

$$S_{MA} = \ell_0(\hat{\beta}_0, \hat{\delta}_0) - \ell(\hat{\gamma}, \hat{\delta}) - \frac{p_0}{2} \log 2\pi + 1/2 \log |I_{\psi \psi}^N(\psi_0, \hat{\beta}, \hat{\delta})| - \log \pi(\psi | \hat{\beta}, \hat{\delta}) \quad (4.3)$$

where $(\hat{\beta}_0, \hat{\delta}_0)$ maximizes the null-hypothetical loglikelihood $\ell_0(\beta, \delta) = \log p(Y | \psi_0, \beta, \delta)$, $(\hat{\psi}, \hat{\beta}, \hat{\delta})$ maximizes the unrestricted loglikelihood $\ell(\psi, \beta, \delta)$, and $I_{\psi \psi}^N(\psi_0, \beta, \delta) = X_\psi' V^{-1}(\delta) X_\psi$ is the portion of the Fisher information matrix corresponding to $\psi$.

Note that $S_{MA}$ is the generalization of $S_A$ for mixed models but in $S_{MA}$, the information term $I_{\psi \psi}^N$ refers to the entire sample. Since the Fisher information matrix and observed information matrix are output of most standard mixed model packages such as SAS (Littell

14
et al. (1996) and BMDP (Dixon et al. (1990)), $S_{MA}$ is still directly computable. The next proposition shows that $S_{MA}$ converges in probability to the Bayes factor using an arbitrary prior $\pi_{\psi|\beta, \delta}(\psi|\beta, \delta)$ for $\psi$ and any choice of regular prior for the nuisance parameters, $(\beta, \delta)$. Its proof is similar to Proposition 3.1 and so is omitted.

**Proposition 4.1** Under Assumption 2 and Conditions 1–4 of Section 3.1, the Bayes factor, $BF_{MA}$, in favor of $M_0 : \psi = \psi_0$ satisfies $\log BF_{MA} - S_{MA} = o_p(1)$.

### 4.3 Default Case

In the default case, where an informative prior, $\pi_{\psi|\beta, \delta}(\psi|\beta, \delta)$, is not available, it may be desirable to use a unit-information prior similar to (3-2) or (3-3) and the subsequent asymptotic approximation to the Bayes factor. For the case where within-subject errors are independent and identically distributed and random effects are a subset of fixed effects, a general approximation may be proposed. The following definition makes precise the type of fixed effects for which it applies.

**Definition 3.1** A fixed effect parameter $\psi_k$ has an associated random effect if its covariate vector $x_{ik}$ is proportional to a column of $Z_i$ for $i = 1, \ldots, I$.

To understand the definition, consider an example where model (4.1) contains only a random intercept for subject. A model of this type is considered in Section 4.4. The random intercept for subject means that $Z_i$ is a vector of ones of length $n_i$ for $i = 1, \ldots, I$. Since between-subject fixed effects are constant for all within-subject observations, their design vectors are proportional to $Z_i$ for $i = 1, \ldots, I$. Within-subject effects, and interactions of within-subject effects with between-subject effects, are not proportional to $Z_i$.

As shown in Section 4.1, the effective sample size, $N_k$, or order of information, for a fixed parameter $\psi_k$ is then determined by whether or not it has an associated random effect. The following definition gives the rule for determining whether $N_k$ is the number of subjects, $I$, or total number of observations, $N$.

**Definition 3.2** $N_k = I$ if $\psi_k$ has an associated random effect and $N$, otherwise.
For testing $M_0 : \psi = \psi_0$ versus $M_1 : \psi \in \mathbb{R}^p$, Definition 3.2 implies that in (2.2), different fixed effects have different effective sample sizes. This suggests use of the criterion

$$S_M = \ell_0(\hat{\delta}_0, \tilde{\delta}) - \ell(\hat{\gamma}, \delta) + \frac{1}{2} \sum_{k=1}^{p_0} \log N_k$$  \hspace{1cm} (4-4)

where $N_k$ is defined by Definition 3.2 for $k = 1, \ldots, p_0$.

Now let $\Delta = \text{diag}(N_1, \ldots, N_{p_0})$ and consider the following prior for $\psi$ under $M_1$:

$$\pi_{\psi|\beta, \delta}(\psi|\beta, \delta) = N_p(\psi_0, \Delta^{1/2}(I_{\psi\psi}^{-1})^{1/2})$$  \hspace{1cm} (4-5)

Prior (4-5) is the analog of (3-2) for mixed models. It follows that $S_M$ converges in probability to the Bayes factor using prior (4-5).

**THEOREM 4.1** In model (4.1) let $\gamma = (\psi, \beta)$ where components of $\psi$ either do or do not have associated random effects as defined by Definition 3.1. Consider a test of $M_0 : \psi = \psi_0$ versus $M_1 : \psi \in \mathbb{R}^p$. Assume the conditions of Proposition 4.1 hold and denote by $B_{NM}$ the Bayes factor in favor of $M_0$ using prior (4.5). Then $\log B_{NM} - S_M = o_P(1)$.

**PROOF:** See Appendix. □

Use of a Cauchy density instead of a Normal density leads to the following modification of $S_M$.

**COROLLARY 4.1** Under the conditions of Theorem 4.1 assume instead that

$$\pi_{\psi|\beta, \delta}(\psi|\beta, \delta) = \text{Ca}_p(\psi_0, \Delta^{1/2}(I_{\psi\psi}^{-1})^{1/2})$$  and let $B_{MC}$ denote the Bayes factor in favor of $M_0$ using this prior. Then $\log B_{MC} - S_{MC} = o_P(1)$ where

$$S_{MC} = S_M + \log r$$  \hspace{1cm} (4-6)

and $r = \pi^{1/2}/\{2^{p_0/2}\Gamma(p_0/2 + 1/2)\}$.

**PROOF:** Similar as in Theorem 4.1 using the Cauchy prior. □

4.4 **EXAMPLE: REPEATED MEASUREMENTS**

In an unpublished study with Dr. Victoria Arango and Dr. Mark Underwood of Columbia University and Dr. Robert Kass of Carnegie Mellon University, we analyzed repeated neuron
area measurements from the locus coeruleus portion of the brain in a group of 31 subjects consisting of 12 controls, 5 suicidal non-alcoholics, 7 non-suicidal alcoholics, and 7 suicidal alcoholics. The aim of the analysis was to see if alcoholics or suicidal patients differed significantly from controls in average neuron area. The data were highly unbalanced; for each subject 35 to 490 neurons were collected at three levels (front, middle and rear) and both sides (right and left) of the locus coeruleus. The total number of observations was 10,269. A random effect for subject was fit to model between-subject variability. The objective then was to choose fixed effects. Between-subject fixed effects of interest were alcohol, suicide, sex, race, age, and postmortem interval. Within-subject effects were side and level of brain. This example illustrates how the Schwarz criterion can be useful for choosing parsimonious models when there is a large amount of data. Relying on standard likelihood ratio tests and p-values for model selection led to models with a large number of seemingly significant effects.

To quickly reduce the large subset of possible models, a collection of $M = 340$ models was fit using BMDP 3V (Dixon et al. (1990)). Since there is a random subject intercept in the model, $N_k = 31$, the number of subjects, for all between-subject effects and 10,269 for all within-subject effects and interactions of between-subject effects with within-subject effects. Approximations (4.4) and (4.6) were computed for each of the $M$ models by comparing them to the simplest model containing only an intercept term. Table 2 shows the top 6 models in terms of highest posterior probabilities using the Schwarz approximations to expressions (2.3). These models comprised roughly 99% of the total probability. Note these probabilities are only crude approximations, of order $O_P(1)$, to the actual probabilities since Schwarz’s increased accuracy holds only for individual pairs of nested hypotheses. Table 2 provides a quick and highly informative summary of model choice with interpretable model probabilities.

We analyzed effects present in each of the models in Table 2. In particular, the effects of alcohol and suicide appeared in all of the models and their parameter estimates implied that neuron area deteriorated in these groups. In a previous analysis we used a naive version of Schwarz, specifying sample size $N = 10,269$ for all effects. This led to a table that placed most of the posterior probability (.89) on Model 6. This model does not contain any of
Table 2: Approximate Posterior Probabilities for Brain Models computed using approximations $S_M$ and $S_{MC}$. Each model had a subject random effect.
the between-subject fixed effects of Model 1. The naive version of Schwarz penalizes these effects too much, using a penalty log (10,269) instead of the appropriate log (32). Our version of Schwarz’s approximation yields a more sensible model with main effects present for interactions.

5 DISCUSSION

We have derived Schwarz’s criterion and two simple modifications for flexible tests of fixed effects in Normal linear mixed models. Under the null hypothesis, $S$ and $S_M$ are approximate Bayes factors using Normal unit-information priors, $S_A$ and $S_{MA}$ allow an arbitrary informative prior, and $S_C$ and $S_{MC}$ correspond to generalizations of the Cauchy prior recommended by Jeffreys (1961). Their simple formula, computable directly from output of most statistical software, and high accuracy, for even small sample sizes, makes the Schwarz criterion and its modifications highly competitive with some of the more computationally intensive exact Bayesian reference procedures. Moreover, it is unclear how some of these latter procedures extend to hierarchical models.

Our arguments justifying Schwarz’s approximation were given in the context of tests of nested hypotheses, under the assumption that the null model or a local alternative is true. An example where this is arguably the situation of interest is the one-way analysis of variance model considered in Section 3-4. A more realistic scenario may be the repeated measures example of Section 4-4, where there is a large collection of models and coherent prior specification is required across all models. Although our asymptotic arguments for Schwarz’s increased accuracy do not hold here, Schwarz is still a crude approximation, adequate for initial model reduction.

Laplace’s approximation is valid for a wide variety of statistical models, including generalized linear models, nonlinear mixed models (Wolfinger (1993) and Vonesh (1996)), and survival type models (Raftery et al. (1994)), implying that Schwarz criteria similar to those discussed in this article are also feasible for these models. We hope to report soon on its application to tests of variance components. Schwarz’s simple formulas provide an attractive addition to more traditional procedures such as the likelihood ratio test.
Proof of Proposition 3.1: Application of Laplace’s method to the numerator and denominator of the Bayes factor, $BF$, leads to the following approximation

$$BF = (2\pi)^{-p_0/2} \left| \frac{\hat{V}_0}{V} \right|^{1/2} \exp \{ \ell_0(\hat{\beta}_0, \hat{\sigma}_0) - \ell(\hat{\psi}, \hat{\beta}, \hat{\sigma}) \} \frac{\pi_0(\hat{\beta}_0, \hat{\sigma}_0)}{\pi(\psi, \beta, \sigma)} (1 + O(n^{-1}))$$

(A-1)

where $\hat{V}_0 = (-D^2 \ell_0(\hat{\beta}_0, \hat{\sigma}_0))^{-1}$, $\hat{V} = (-D^2 \ell(\hat{\psi}, \hat{\beta}, \hat{\sigma}))^{-1}$, and $\cdot \cdot$ denotes determinant.

The conditions stated satisfy the analytical assumptions for Laplace’s method (Kass et al. (1990), pg 476, 479). Assumption 1 is sufficient to guarantee the family of linear model densities is Laplace regular and hence for the approximation to hold with probability one (Kass et al. (1990), pg 484).

For the remainder of the proof suppose that $\psi$ and $\beta$ are one-dimensional. The multi-dimensional case is proven in the same manner. Let $(\hat{\beta}_0, \hat{\sigma}_0)$ be the maximum likelihood estimate of $\ell(\psi_0, \beta, \sigma)$. Since $\partial_\beta \ell(\hat{\psi}, \hat{\beta}, \hat{\sigma}) = 0$ when $\partial_\beta \ell(\psi_0, \hat{\beta}_0, \hat{\sigma}_0)$ is expanded about $(\psi, \beta, \sigma)$ we obtain:

$$0 = \partial_\beta \ell(\psi_0, \hat{\beta}_0, \hat{\sigma}_0) = (\psi_0 - \hat{\psi}) \partial_{\psi \psi}^2 \ell(\hat{\psi}, \hat{\beta}, \hat{\sigma}) + (\hat{\beta}_0 - \hat{\beta}) \partial_{\beta \beta}^2 \ell(\hat{\psi}, \hat{\beta}, \hat{\sigma})$$

$$+ (\hat{\sigma}_0 - \hat{\sigma}) \partial_{\sigma \sigma}^2 \ell(\hat{\psi}, \hat{\beta}, \hat{\sigma}) + O_P(1).$$

For linear models $\partial_{\beta \sigma}^2 \ell(\hat{\psi}, \hat{\beta}, \hat{\sigma}) = 0$ and $X_{n0}$ orthogonal to $X_{n1}$ implies $\partial_{\psi \psi}^2 \ell(\hat{\psi}, \hat{\beta}, \hat{\sigma}) = 0$. Since $\partial_{\beta \beta}^2 \ell(\hat{\psi}, \hat{\beta}, \hat{\sigma}) = O_P(n)$, we obtain $\hat{\beta}_0 - \hat{\beta} = O_P(n^{-1})$. Replacing the first $\beta$ by $\sigma$ in each subscript of the above expression and noting that $\partial_{\psi \psi}^2 \ell(\hat{\psi}, \hat{\beta}, \hat{\sigma}) = 0$ and $\partial_{\beta \beta}^2 \ell(\hat{\psi}, \hat{\beta}, \hat{\sigma}) = O_P(n)$ yields $\hat{\sigma}_0 - \hat{\sigma} = O_P(n^{-1})$. This along with $\hat{\psi} - \psi_0 = O_P(n^{-1/2})$, orthogonality, and $-\frac{1}{n} D^2 \ell(\hat{\psi}, \hat{\beta}, \hat{\sigma}) - I(\psi, \beta, \sigma) = O_P(n^{-1/2})$, where $I(\psi, \beta, \sigma)$ denotes the Fisher information in a sample of size one, implies:

$$| - D^2 \ell_0(\hat{\beta}_0, \hat{\sigma}_0) |^{-1/2} | - D^2 \ell(\hat{\psi}, \hat{\beta}, \hat{\sigma}) |^{1/2} = n^{p_0/2} | I_{\psi \psi}(\psi_0, \hat{\beta}, \hat{\sigma}) |^{-1/2} (1 + O_P(n^{-1/2})).$$

(A-2)
Finally, Condition 2, \( \hat{\beta}_0 - \hat{\beta} = O_P(n^{-1}) \) and \( \hat{\sigma}_0 - \sigma = O_P(n^{-1}) \) imply

\[
\frac{\pi_0(\hat{\beta}_0, \hat{\sigma}_0)}{\pi(\hat{\beta}, \sigma)} = \pi_{\psi|\beta, \sigma}(\hat{\beta}, \sigma)^{-1}\{1 + O_P(n^{-1})\}. \tag{A-3}
\]

Substituting (A-2) and (A-3) into (A-1) yields (3.1). \( \square \)

Proof of Result 3.1: We give the formulas for the log of the Bayes factor in favor of the null hypothesis for Schwarz (SC), the method of Zellner and Siow (ZS), and the fractional Bayes factor (FBF). The formula for SC is given below (3.3), the method for calculating ZS in Zellner & Siow (1980), and the method for FBF in O’Hagan (1995).

\[
SC = -\frac{n}{2} \log r + \frac{p_0}{2} \log n + \log c_s,
\]

\[
ZS = SC + \frac{p_0 + p_1 + 1}{2} \log r,
\]

\[
FBF = -\frac{n - p_0 - p_1 - 1}{2} \log r + \log c_f,
\]

where \( c_s = \Gamma(1/2)/\{2^{p_0/2}\Gamma(p_0/2 + 1/2)\} \), \( c_f = \Gamma(n/2 - p_1/2)\Gamma(1/2)/\{\Gamma(n/2 - p_0/2 - p_1/2)\Gamma(p_0/2 + 1/2)\} \), \( r \) equals \( (s_A^2/s_0^2)^{-1} \) where \( s_0^2 \) is the residual sum of squares under the null and \( s_A^2 = s_0^2 - \hat{\psi}\hat{X}'_n\hat{Y}_n\hat{Y}_n\hat{\psi} \) is the residual sum of squares under the alternative. Note that \( r = \{1 - O(n^{-1})\}^{-1} \). Now using Stirling’s expansion, \( \Gamma(x/2) = (2\pi)^{1/2}(x/2)^{x/2-1/2}\exp(-x/2){1+O(1/x)} \), and Taylor’s expansion, \( \log(x+1) = x + O(x^2) \), yields the result.

The following Lemma will be necessary for the proof of Theorem 4.1.

**Lemma 5.1** Let \( Z = [z_1 : Z_0] \) be a \( n \times q \) design matrix where \( z_1 \) is \( n \times 1 \) and \( Z_0 \) is \( n \times q - 1 \). Suppose the \( q \times q \) matrix D for Z is \( D = \text{diag}(\sigma_1^2, D_0) \) where \( \sigma_1^2 \) is the variance component corresponding to \( z_1 \) and \( D_0 \) is the \((q - 1) \times (q - 1)\) variance matrix corresponding to \( Z_0 \). Then

\[
\frac{1}{\sigma^2} z_1'[I_n - ZD(\sigma^2 I_q + Z'D)^{-1}Z']z_1 = (z_1'[\sigma_1^2 + \sigma^2 - z_1'Z_0D_0(Z_0'Z_0D_0 + \sigma^2 I_{q-1})^{-1}Z_0'z_1]\sigma_1^2)^{-1} \times (z_1'[z_1 - z_1'Z_0D_0(Z_0'Z_0D_0 + \sigma^2 I_{q-1})^{-1}Z_0'z_1]) \tag{A-4}
\]

**Proof:** *Directly from algebra.* \( \square \)
PROOF OF THEOREM 4.1:

Consider the diagonal elements of the information matrix given by (4-2). If \( D = \text{diag}(\sigma_1^2, D_0) \), the expression for the \( i \)th element in (4-2) is given by (A-4) with \( z_i = x_{ik} \), \( Z = Z_i \), and \( n = n_i \) for \( i = 1, \ldots, I \). If \( \sigma_1^2 > 0 \) expression (A-4) is \( O_P(1) \) and hence the order of (4-2) is \( I \). If \( \sigma_1^2 = 0 \) the order of (A-4) is \( n_i \) and hence the order of (4-2) is \( N \). The case \( \sigma_1^2 > 0 \) corresponds to the case where \( \psi_k \) has a random component, as defined by Definition 3.1, and the case \( \sigma_1^2 = 0 \) corresponds to the case where \( \psi_k \) has no random component. The nondiagonal \( D \) case follows similarly. From this it follows that \( \lim_{I, n_i \to \infty} \Delta^{-1/2} I_{\psi \psi}^N \Delta^{-1/2} \) exists and is positive definite.

In Proposition 4.1 substitute

\[
\pi_{\psi|\beta, \delta}(\hat{\beta}, \hat{\delta})^{-1} = (2\pi)^{p_\psi/2} |\Delta^{1/2}(I_{\psi \psi}^N)^{-1}\Delta^{1/2}|^{1/2} \exp\left\{ \frac{1}{2} \psi' (\Delta^{1/2}(I_{\psi \psi}^N)^{-1}\Delta^{1/2})^{-1} \psi \right\} \\
= (2\pi)^{p_\psi/2} |\Delta|^{1/2} |I_{\psi \psi}^N|^{-1/2} \{1 + O_P(1)\}
\]

under \( M_0 \), \( \square \).

ACKNOWLEDGEMENT

This research is part of the author’s doctoral dissertation completed under the supervision of Dr. Robert E. Kass at Carnegie Mellon University, and funded by a National Science Foundation graduate fellowship. The author gratefully acknowledges the help of Dr. Kass and the constructive comments of an anonymous referee which greatly improved the final manuscript.

REFERENCES


