This is to certify that the

dissertation entitled

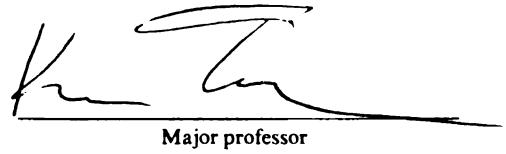
A COMPARISON OF ALTERNATIVE APPROXIMATIONS TO MAXIMUM LIKELEHOOD ESTIMATION FOR HIERARCHICAL GENERALIZED LINEAR MODELS: THE LOGISTIC-NORMAL MODEL CASE

presented by

Matheos Yosef

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in CEPSE



Major professor

Date 8-27-01

A COMPARISON OF ALTERNATIVE APPROXIMATIONS TO MAXIMUM LIKELIHOOD ESTIMATION FOR HIERARCHICAL GENERALIZED LINEAR MODELS: THE LOGISTIC-NORMAL MODEL CASE

By

Matheos Yosef

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational Psychology and Special Education

2001

ABSTRACT

A COMPARISON OF ALTERNATIVE APPROXIMATIONS TO MAXIMUM LIKELIHOOD ESTIMATION FOR HIERARCHICAL GENERALIZED LINEAR MODELS: THE LOGISTIC-NORMAL MODEL CASE

By

Matheos Yosef

Educational data often have hierarchical structure (e.g., students are nested within clusters such as schools). Also, outcome variables can sometimes be discrete (e.g., whether a student repeats a grade). In such cases, the outcome variable is usually related to the covariates and cluster random effects using a hierarchical generalized linear model.

The (marginal) maximum likelihood (ML) estimation method is widely used to estimate the parameters of such models. To obtain the marginal likelihood formula that needs to be maximized, the random effects must be integrated out of the joint distribution of the outcome and the random effects. In many cases, the integration cannot be carried out in closed form. Several approaches have been used to approximate this integral. This dissertation compared four methods of integral approximation -- two Laplace-based and two based on Gaussian numerical integration.

Analytic and numerical comparisons show that, for the univariate random effects model case, the 2nd order Laplace method (Laplace2) and the adaptive Gauss-Hermite method (AGH) with one quadrature point give the same result. The 6th order Laplace approximation method (Laplace6) has the same order of error as the AGH with 4 (to 6) quadrature points. It took much more (8 and 14) quadrature points for the ordinary Gauss-Hermite (GH) to give results similar to Laplace2 and Laplace6. The error of Laplace6

approximation was better than that of Laplace2 by at least $O(n^{-1})$, where n is the cluster size.

Simulation studies using programs (HLM, MIXOR and SAS PROC NLMIXED) that implement the four methods indicate that, for a univariate random effects case, all methods perform well when the cluster size is quite large. However, Laplace2 usually gives the most biased estimates and sometimes has the largest mean-squared errors (MSE). For a small cluster size, AGH performed the best as far as speed, MSE and bias are concerned, while the ordinary GH performed the worst. For a multivariate (bivariate) random effects model case, Laplace6 performed the best (in terms of bias and MSE) with the ordinary GH following closely. The estimates of Laplace2 had the largest biases while the algorithm implementing AGH was computationally the slowest and needed specification of good starting parameter values.

Overall, Laplace2 appears to be a simple and fast method to get estimates, especially for a model with small random effects variance. Laplace6 is much more accurate and quite fast but needs derivation of cumbersome formulas. GH is quite simple but may need a fairly large number of quadrature points for an accurate estimation. This makes it computationally inefficient for a multivariate random effects case. AGH combines the advantages of GH (simplicity of formula) and high-order Laplace (accuracy with quite few quadrature points) but it can also be computationally quite inefficient as the dimension of random effects increases. The Laplace-based methods do not suffer from dimensionality problem that much.

To my parents

ACKNOWLEDGMENTS

First and foremost, I would like to give thanks to my Lord and Savior Jesus Christ for the grace He has given me.

I am indebted to all my committee members for their comments, suggestions, questions and challenges that motivated the work in this dissertation. My advisor, Dr. Stephen Raudenbush, deserves my special thanks for his advice, guidance, instruction and support during my doctoral program. My deep gratitude is also due to my committee chairperson, Dr. Ken Frank, for facilitating my dissertation process as well as for his insightful suggestions and comments. I would also like to thank my other committee members, Dr. Betsy Becker and Dr. Habib Salehi, for their comments, corrections, questions as well as encouragement and support.

I am grateful to my family as well as my prayer partners and fellowship members, both in Lansing and Ann Arbor, for their prayers, encouragement and support. Thanks are also due to my colleague and former co-student, Dr. Meng-Li Yang, for her motivation, encouragement and prodding, as well as to my colleagues and friends, Dr. Yuk Fai Cheong, Dr. Yasuo Miyazaki and Christopher Johnson, for their comments, suggestions and encouragement. Finally, I would like to thank Xiang Gui for his help in the proof of one asymptotic result (that of the adaptive Gauss-Hermite).

To God be all the glory.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Education is important to society. Many are concerned with the quality of education that students are getting. With such problems as drop-out and juvenile delinquency, they may also be interested in knowing what factors affect and facilitate student learning, and what relationship educational factors have to social issues such as crime. Educational studies are conducted to address such issues. Some studies involve large-scale quantitative educational data with measurements on students as well as some aspects of the educational system such as schools and teachers.

Study participants (e.g., students) are often observed within a certain context or "cluster." For example, students are nested within classes, classes within schools, and schools within school districts. We are not only interested in student characteristics but also such contextual effects as effects of teacher and school characteristics on student outcomes.

In other cases, subjects are repeatedly observed over time to monitor or measure growth. In such repeated measures data, occasions of observations are nested within individuals. Data that have such nested structures, whether they are the cross-sectional clustered data or longitudinal repeated measures data, are known as hierarchical (e.g., Bryk and Raudenbush, 1992) or multilevel (e.g., Goldstein, 1987) data.

In recent times, educational researchers have adopted hierarchical models to analyze hierarchical educational data. Hierarchical models enable researchers to ascertain the interactions between the different levels (students, teachers, schools, policies, etc) as well

as to discern the amount of variation explained at each level. Applications of hierarchical linear models, variously known as linear mixed models (Goldstein, 1986) or random coefficient models (Rosenberg, 1973; Longford, 1993) or covariance components models (Dempster, Rubin and Tsutakawa, 1981), include relations of school effectiveness to student achievement scores (Raudenbush and Bryk, 1986; Aitkin and Longford, 1986; Young, 1996), effects of teacher interaction outside the classroom on student learning (Louis et al., 1994), effects of the races of rater as well as ratee on evaluations of performance (Waldman and Avolio, 1991). Detailed descriptions of the models, their methodology as well as applications in education and/or social sciences are given by Goldstein (1995), Bryk and Raudenbush (1992), Longford (1993), Bock (1989) and Raudenbush and Willms (1991).

Hierarchical models take into account the dependence that usually exists between observations in the same cluster such as the school. Owing to cluster effects, observations in the same cluster cannot be expected to be independent. Regression models that assume independence between observations are thus often misspecified. For instance, Aitkin et al. (1981) found no difference between 'formal' and non-formal teaching when analyzing the student data nested within classes, while Bennet (1976), who used ordinary multiple regression on the same data without nesting (grouping students into classes), had found a difference. A mixed-effects regression model, which includes both the cluster random effects as well as the fixed effects of the covariates on the outcome, is generally used to model such hierarchical data.

Also, outcome variables are sometimes discrete (rather than continuous). In educational studies, the outcome can be grade retention or dropping out of school (Rumberger, 1995). Horney, Osgood and Marshall (1995) analyzed the effects of local life circumstances, including school, on the likelihood of committing felonies. In these cases, the outcome is dichotomous and usually modeled using a binomial distribution. We might also want to study what student, school or neighborhood factors affect the number of days a student is absent, or what factors determine the number of felonies committed by a student in a school in a given time. This kind of outcome, a count, is usually modeled using a Poisson distribution. Using hierarchical linear model analysis, Bryk and Thum (1989) found that "high levels of internal differentiation (i.e., among students) within high schools and weak normative environments contribute to the problems of absenteeism and dropping out" while "no single factor makes schools effective in sustaining student interest and commitment." Rumberger (1995) used hierarchical linear modeling to study "dropouts from middle school and examine the issue from both individual and institutional (school) perspectives."

Just as hierarchical linear models are used to analyze hierarchical data with normal outcome variables, hierarchical generalized linear models, which are generalized linear models with random effects (McCullagh and Nelder, 1989), are used to model and analyze hierarchical data with non-normal outcomes such as binary and count data. Hierarchical generalized linear models (HGLMs) are also known as generalized linear mixed models (GLMMs) (e.g., Breslow and Clayton, 1993). At the first level of the hierarchy ("level 1"), a generalized linear model is substituted for the linear regression

model of the normal data. The coefficients of this model then vary over clusters at a second level ("level 2"). At this second level, linear models predict these level 1 coefficients using cluster characteristics as explanatory variables. Random effects at level 2 model the unexplained variation in the level 1 coefficients. The resulting combined model is a generalized linear model with random effects. Breslow and Clayton (1993) describe such models and provide an approach to estimating them, giving examples of dichotomous and count outcomes. Stiratelli, Laird and Ware (1984) had already described a random-effects model with binary (dichotomous) response and provided an estimation procedure.

In order to estimate the parameters (i.e., the fixed effects and the variance components of the random effects) of such a model, recourse is usually made to (marginal) maximum likelihood (ML) estimation method because ML estimates have such well-known, large sample properties as consistency, asymptotic normality and efficiency (minimum variance). This method maximizes the likelihood of the observed outcome data. In hierarchical models, we may specify the conditional distribution of the outcome variable y, given the random effect b as $f(y|b)$. In order to find the marginal likelihood of the outcome, the conditional likelihood is mixed with the density of the random effect, $p(b)$. That is to say, the conditional distribution of the outcome variable is multiplied by the density of the random effect, and the random effect is integrated out leaving the marginal likelihood of the outcome, viz, $h(y)=\int f(y|b)p(b)db$. Unless the density of the random effect is the conjugate prior for the conditional distribution of the outcome, a closed form formula cannot generally be found for the marginal likelihood of the outcome.

The multivariate normal prior, along with the multivariate t, are well-suited to modeling correlated random effects per cluster (Raudenbush, 1999). Thus, the density of the random cluster effect, $p(b)$, is often assumed to be normal. However, since this density of the random effect is not the conjugate prior for the conditional distributions of such non-normal outcomes as binary and count data (their conditional distributions are usually taken to be binomial and Poisson, respectively) , the integration cannot usually be carried out analytically to obtain the marginal likelihood. So, in the cases of non-normal data at level 1 along with normal random effects, we may not have the marginal likelihood of the outcome in a closed form (see Zeger et al., 1988, in the case of the hierarchical logistic model).

In this dissertation, several approaches to approximate (marginal) maximum likelihood for the hierarchical generalized linear model will be compared as to performance. Specifically, I will compare approaches to estimating the hierarchical logistic model. To this end, first, formulas for the various approaches to approximate the integral are derived. The approaches used (and compared) in this dissertation are the standard Laplace, a 6th order Laplace, and non-adaptive and *adaptive* Gauss-Hermite quadratures. The formulas are compared analytically in simple cases to see which approaches fare better. Graphs are used for illustration.

Next, datasets with different models and parameter values are simulated and analyzed using the different approaches and the results compared for performance (accuracy and efficiency) of the approaches. For Laplace, the HLM program implementing Raudenbush's posterior modal algorithm (1992) is used, while its Higher-order Laplace

option (Raudenbush, Yang and Yosef, 2000) is used for the 6th order Laplace. Hedeker and Gibbons' (1994) MIXOR program is used for the non-adaptive Gaussian quadratures estimation while the *adaptive* Gaussian estimation is carried out using the SAS procedure PROC NLMIXED (Wolfinger, 1999) for the estimation of non-linear mixed models which has *adaptive* Gaussian quadratures as the default option.

Finally, the various methods are used to analyze the large scale educational dataset of the 1988 National Survey of Primary Education in Thailand (Thailand data). The results from the different methods are compared, mostly for similarity to each other or divergence from each other, in light of the results from the analytic and simulation-based comparisons.

Chapter 2

BACKGROUND AND SIGNIFICANCE

One popular method used to estimate the parameters of generalized linear mixed-model (GLMM) is the maximum likelihood (ML) method. This method finds parameter estimates that maximize the likelihood of the observed data. In order to do this, the likelihood of the data must first be obtained. In a GLMM case, we obtain the likelihood of the data by integrating out the random effects from the joint distribution (likelihood) of the data and random effects. Oftentimes, the integration cannot be carried out in closed form. One such model that is fairly common in various fields including education is the logit-normal mixed model. In this case, conditional on the random effects, the data have a binomial distribution with a conditional mean that is related to the linear predictor (the sum of the fixed and random effects) via a canonical link function (McCullagh and Nelder, 1989), while the random effects are assumed to have a multivariate normal distribution. Researchers have used various approaches to approximate the likelihood that must be maximized. A brief and general review of the approaches follow under two broad headings.

Quasi-likelihood and Approximate Likelihood Approaches

Longford (1993, 1994) approximated the marginal likelihood by taking a second order Taylor series expansion of the joint likelihood around zero (i.e., b=0) and then making use of normal theory to integrate the approximation. This result turned out to be the same as Goldstein's (1991) iterative generalized least squares (IGLS) approach which used the linearized dependent variable (McCullagh and Nelder, 1989) transforming the (discrete)

outcome into a continuous one. Goldstein didn't assume normality but the existence of the first two moments. This method was labeled marginal quasi-likelihood (MQL) by Breslow and Clayton (1993) because it involves expanding the conditional expectation around zero for the random effects. Rodriguez and Goldman (1995) evaluated two packages based on MQL and found out that MQL estimates of both the fixed effects and the variance components exhibit downward biases (biases toward zero), and the biases are more pronounced when the random effects have large variances.

Breslow and Clayton (1993) applied Laplace's method for integral approximation to approximate the quasi-likelihood function, which is equivalent to the true likelihood if conditionally on the random effects the observations are drawn from a linear exponential family (as in the logit-normal mixed model case). The Laplace method expands the exponent of the integrand, Green's (1987) penalized quasi-likelihood (PQL), expressed as a function of the random effects, in a second-order Taylor series around the maximizer (known as the conditional mode) of the exponent function and uses normal theory to find the integral. For the canonical link functions, the log of the approximate integral (quasi-likelihood) turns out to be Green's (1987) PQL evaluated at the conditional mode plus a function of the covariance matrix of the random effects and the GLM iterated weights (McCullagh and Nelder, 1989). Assuming the GLM iterative weights vary slowly as a function of the mean, they maximized only Green's PQL for the fixed effects and random effects using Fisher scoring and normal theory, and used pseudo-likelihood for the variance components estimation. The score equations they used to maximize for the fixed and random effects were also derived by Stiratelli, Laird and Ware (1984) for logistic

regression of binary data by maximizing the posterior distribution for the fixed and random effects under a diffuse prior for the fixed effects.

Raudenbush (1992) extended Stiratelli et al.'s (1984) joint posterior modal approach for binary outcomes -- where inferences were based on the joint posterior modes of the regression coefficients given approximate REML covariance estimates -- to a broad class of hierarchical generalized linear models and used Schall's (1991) framework to improve the efficiency of his approach. Although this PQL approach improves upon the MQL approach as far as parameter estimation in the hierarchical model is concerned, PQL estimators of the variance components were still subject to serious bias when applied to correlated binary data with large variances (Yang, 1994; Breslow and Lin, 1995). The latter provided a correction to the bias via fourth-order expansion of the joint distribution of the data and the random effects around the current estimates, followed by Laplace's method to approximate the marginal likelihood.

Yang (1998) extended the expansion of the exponent of the integrand (the joint likelihood of the data and random effects) to the sixth order Taylor's series and then used normal theory to do the integration obtaining approximate likelihood function. She did this for the multiple random effects case with a general variance-covariance matrix. She used the output from Raudenbush's posterior modal algorithm (1992) as starting values for the parameters in order to ensure convergence and more efficient estimation. She then used Fisher scoring to simultaneously estimate the fixed effects and the variance-covariance components of the random effects. Her method was a big improvement over PQL in terms of results as well as being computationally quite efficient. She even went

on to eighth order Laplace approximation but the results didn't improve so she settled for the sixth order. Nevertheless, she provided an infinite multivariate Taylor series expansion that would virtually make the approximate marginal likelihood equivalent to an actual likelihood.

Full Likelihood Approach

Anderson and Aitkin (1985) used Gaussian quadrature formulas with the logistic model with a single random effect to integrate (approximately) the joint likelihood of the data and the random effect, with respect to the random effect, in order to find the (approximate) marginal likelihood of the data. Hedeker and Gibbons (1994, 1996) extended this to the probit and logistic models with multiple random effects and used Fisher scoring to maximize the resulting (approximate) marginal likelihood.

With this numerical Gaussian quadrature integration formula, the approximation to the marginal likelihood gets better as the number of quadrature points increases. However, as the dimension of the random effects increases, an increase in the number of quadrature points in one dimension increases exponentially the total number of quadrature points (and hence computations) required for all the random effects. This exponential computational complexity in the case of random effects is the major drawback of the Gaussian quadrature procedure. Bock, Gibbons and Muraki (1988), however, noted that the number of quadrature points for each dimension (random effect) can be reduced, without appreciably harming the approximation, as the number of dimensions increases.

Yosef (1997) conducted a simulation study of a two-level mixed-effects logit model with a single random effect using Gauss-Hermite formulas as implemented in Hedeker and Gibbons' MIXOR program and found out that, in general, it gives better estimates than PQL in terms of biases as well as comparable ones in terms of mean square errors. However, the simulation study also revealed that in some instances, especially when the random effects variance and the average probability of success are small, MIXOR either gave unreasonable estimates or no estimates at all while PQL almost always gave reasonable estimates.

For the nonlinear mixed-effects model, where a continuous outcome is a nonlinear function of the fixed and random effects, Pinheiro and Bates (1995) argued that Gaussian quadrature centered at the expected value of the random effects is quite inaccurate for a smaller number of abscissas (quadrature points) and computationally inefficient for a larger number of abscissas. So, they first expanded the exponent of the integrand (the joint likelihood of the data and random effects expressed as a function of the random effects) in a second order Taylor series expansion around the conditional mode of the random effects just as was done in PQL. This has the effect of centering the joint likelihood around the conditional mode rather than the random effects mean of 0. Then, they applied Gaussian quadrature on the resulting integrand to approximate the (marginal) likelihood which they maximized to find the parameter estimates. They labeled their procedure *adaptive* Gaussian quadrature and found it to be quite accurate and computationally efficient which is achieved by a reduction in the number of quadrature points needed. Liu (1993) and Liu and Pierce (1994) also gave the formula for the

adaptive (though they didn't call it so) Gauss-Hermite and used it in examples to obtain likelihoods. They also worked out the asymptotic behavior of the adaptive Gauss-Hermite. Wolfinger (1999) implemented the adaptive Gaussian procedure for a broad class of mixed models including GLMMs in SAS.

McCulloch (1997) developed three algorithms for maximum likelihood (ML) in GLMMs. He constructed a Monte Carlo version of the EM (MCEM) algorithm for GLMMs by incorporating a Metropolis-Hastings step. He also proposed a Monte Carlo Newton-Raphson (MCNR) procedure and evaluated and improved on the simulated ML (SML) method which uses importance sampling. Whereas the first two, MCEM and MCNR, work on the log of the likelihood, the third one estimates the likelihood directly using importance sampling. He also suggested a hybrid algorithm with a preliminary stage of MCEM or MCNR followed by SML. He found his methods to perform better than joint maximization methods such as PQL. However, as is well known, Monte Carlo methods are computationally intensive (time consuming) and convergence is stochastic. He noted that for sufficiently large simulation sample sizes, the Monte Carlo versions would inherit the properties of the exact versions: MCEM would, under suitable regularity conditions, converge to a local maximum whereas NR algorithms are not guaranteed convergence when the surfaces to be maximized are not concave. In view of the stochastic convergence (getting "close" to the correct answer and then varying in that neighborhood), he commented that "this is one reason for suggesting a follow-up round of SML, to avoid the complications of deciding whether the stochastic versions of EM or NR have converged."

Chapter 3

APPROXIMATING THE LIKELIHOOD : AN ILLUSTRATIVE EXAMPLE

Introduction

In this chapter, I will show through a simple example how the methods under study approximate the likelihoods of interest. Let us assume we have a binary outcome variable *Y*. Consider the simple mixed logistic model:

$$\eta_i = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta + b, \quad b \sim N(0,\tau) \ , \tag{3.1}$$

where the intercept β is the fixed effect and b the random effect for a specific cluster and $\mu_i = P(Y_i = 1|b)$ is the conditional probability of success for the binary outcome variable Y given the random effect b. We are interested in finding the maximum likelihood estimates of the parameters in the model (3.1), namely β and τ. From (3.1), we have

$$\mu_i = \frac{1}{1+\exp[-\eta_i]} = \frac{1}{1+\exp[-(\beta+b)]} \Rightarrow \frac{\partial \mu_i}{\partial \eta_i} = \mu_i(1-\mu_i) \equiv w_i \ . \tag{3.2}$$

The likelihood function of the observed data y is given by

$$\begin{aligned} L(\beta,\tau;y) = f_y(y) &= \int_{-\infty}^{\infty} f_{y|b}(y|b)g(b)db \\ &= (2\pi\tau)^{-1/2} \int_{-\infty}^{\infty} \prod_{i=1}^{n} \mu_i^{y_i}(1-\mu_i)^{1-y_i} \exp(-\frac{1}{2}b^2/\tau)db \\ &= (2\pi\tau)^{-1/2} \int_{-\infty}^{\infty} \exp(h(b))db \ , \end{aligned} \tag{3.3}$$

where

$$h(b)=\sum_{i=1}^{n}[y_i\log(\mu_i)+(1-y_i)\log(1-\mu_i)]-\frac{1}{2}b^2\tau^{-1}$$
$$=\sum_{i=1}^{n}[y_i\log\left(\frac{\mu_i}{1-\mu_i}\right)+\log(1-\mu_i)]-\frac{1}{2}b^2\tau^{-1}\ . \tag{3.4}$$

Now, there is no closed-form solution to the integral (3.3). So, we resort to integral approximations (in this dissertation, Laplace and Gaussian based approximations). The goal is to approximate the integrand (as well as the integral) as closely as possible. Since both the Laplace approaches as well as the adaptive Gauss-Hermite require the first and second derivatives of h with respect to b, I will give them here before proceeding. Thus, we have

$$h'(b)=\sum_{i=1}^{n}[y_i\frac{\partial\eta_i}{\partial b}+\frac{1}{1-\mu_i}(-\frac{\partial\mu_i}{\partial b})]-\frac{b}{\tau}$$
$$=\sum_{i=1}^{n}[y_i\frac{\partial\eta_i}{\partial b}+\frac{1}{1-\mu_i}(-\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial b})]-\frac{b}{\tau} \tag{3.5}$$
$$=\sum_{i=1}^{n}(y_i-\mu_i)-\frac{b}{\tau}\ =\ \sum_{i=1}^{n}\left(y_i-\frac{1}{1+\exp[-(\beta+b)]}\right)-\frac{b}{\tau}$$

and

$$h''(b)=\sum_{i=1}^{n}(\frac{\partial\mu_i}{\partial b})-\frac{1}{\tau}=-[\sum_{i=1}^{n}\mu_i(1-\mu_i)+\frac{1}{\tau}]. \tag{3.6}$$

The Laplace approaches and the adaptive Gauss-Hermite involve expanding h around its maximizer $\hat{b}$ (also known as the conditional mode). To find the $\hat{b}$ that maximizes $h(b)$,

we set

$$h'(b)=0 \Leftrightarrow \frac{\hat{b}}{\tau}+\frac{n}{1+\exp[-(\beta+\hat{b})]}=\sum_{i=1}^{n} y_i. \tag{3.7}$$

Illustrative Example

Let $n_j=n=10$ and $J=1$ (i.e., one cluster). Also, let $\tau=1$, $\beta=-1$, and the observed data vector $y=(1,0,0,0,1,1,0,1,0,0)'$. Then, equation (3.7) becomes

$$\hat{b}+\frac{10}{1+\exp[-(-1+\hat{b})]}=4 \Rightarrow \hat{b}=.41717. \tag{3.8}$$

The integral in (3.3) was computed to be .001473319 using the Trapezoidal Rule (Mathews, 1987) with error less than 10^{-8} and limits of integration (-5,5). This was taken to be the actual (true) value of the integral. I now proceed to show how well the various approaches I am considering here approximate the integrand as well as the integral. I will use graphs to further illustrate the integrand approximations.

Laplace Approximations

Using the Taylor-series expansion of h around its maximizer $\hat{b}$, the integral in (3.3) can be written as

$$\begin{aligned}\int_{-\infty}^{\infty}\exp[h(b)]db &= \exp[h(\hat{b})]\int_{-\infty}^{\infty}\exp[-\frac{1}{2}\frac{(b-\hat{b})^2}{V}]\exp[S]db \\ &= (2\pi V)^{1/2}\exp[h(\hat{b})]E(\exp[S])\end{aligned} \tag{3.9}$$

where

$$V=-[h''(\hat{b})]^{-1},$$
$$S=\sum_{k=3}^{\infty} T_k \ , \quad and \quad T_k=h^{(k)}(\hat{b})\frac{(b-\hat{b})^k}{k!} \ . \tag{3.10}$$

The Taylor-series expansion of the exponential function about 0 implies

$$E(e^{S})=E(1+S+\frac{1}{2}S^2+...). \tag{3.11}$$

Note that $E(T_k)=0$ for odd k since the expectation is taken over $N(0,V)$. The first few Laplace approximations (to the integral (3.9)) are defined by approximating $E(e^S)$ as follows:

$$\begin{aligned}
&Laplace2\text{: } E(e^{S})\approx 1\\
&Laplace4\text{: } E(e^{S})\approx 1+E(T_4)\\
&Laplace6\text{: } E(e^{S})\approx 1+E(T_4)+E(T_6)+\frac{1}{2}E(T_3^2)\\
&Laplace8\text{: } E(e^{S})\approx 1+E(T_4)+E(T_6)+E(T_8)+\frac{1}{2}\{E(T_3^2)+E(T_4^2)+2E(T_3T_5)\}
\end{aligned} \tag{3.12}$$

where (see Raudenbush, Yang and Yosef, 2000)

$$\begin{aligned}
&T_3=-\sum_{i=1}^{n}\hat{w}_i(1-2\hat{\mu}_i)\frac{(b-\hat{b})^3}{3!}=-n\hat{w}_i(1-2\hat{\mu}_i)\frac{(b-\hat{b})^3}{3!};\ \hat{\mu}_i=\frac{1}{1+\exp[\beta+\hat{b}]};\ \hat{w}_i=\hat{\mu}_i(1-\hat{\mu}_i)\ ,\\
&T_4=-\sum_{i=1}^{n}\hat{w}_i(1-6\hat{w}_i)\frac{(b-\hat{b})^4}{4!}=-n\hat{w}_i(1-6\hat{w}_i)\frac{(b-\hat{b})^4}{4!}\ ,\\
&T_5=-n\hat{w}_i(1-2\hat{\mu}_i)(1-12\hat{w}_i)\frac{(b-\hat{b})^5}{5!}\ ,\\
&T_6=-n\{\hat{w}_i(1-6\hat{w}_i)(1-12\hat{w}_i)-12\hat{w}_i^2(1-2\hat{\mu}_i)^2\}\frac{(b-\hat{b})^6}{6!}\ ,\\
&T_8=-n\hat{w}_i(1-126\hat{w}_i+1680\hat{w}_i^2-5040\hat{w}_i^3)\frac{(b-\hat{b})^8}{8!}\ ,
\end{aligned} \tag{3.13}$$

whence

$$
\begin{aligned}
E(T_4) &= -n\hat{w}_i(1-6\hat{w}_i)\frac{E(b-\hat{b})^4}{4!} = -n\hat{w}_i(1-6\hat{w}_i)\frac{3V^2}{24} = -n\hat{w}_i(1-6\hat{w}_i)\frac{V^2}{8} \ , \\
E(T_6) &= -n\{\hat{w}_i(1-6\hat{w}_i)(1-12\hat{w}_i) - 12\hat{w}_i^2(1-2\hat{\mu}_i)^2\}\frac{V^3}{48} \ , \\
E(T_3^2) &= \frac{5}{12}V^3[n\hat{w}_i(1-2\hat{\mu}_i)]^2 \ , \\
E(T_8) &= -n\hat{w}_i(1-126\hat{w}_i+1680\hat{w}_i^2-5040\hat{w}_i^3)\frac{V^4}{384} \ , \\
E(T_4^2) &= \frac{35}{192}V^4[n\hat{w}_i(1-6\hat{w}_i)]^2 \ , \\
E(T_3T_5) &= \frac{7}{48}V^4n^2\hat{w}_i^2(1-2\hat{\mu}_i)^2(1-12\hat{w}_i) \ .
\end{aligned}
\tag{3.14}
$$

Similarly, the *integrand* in (3.3) (or the left-hand integral in (3.9)) is approximated using the Laplace technique as

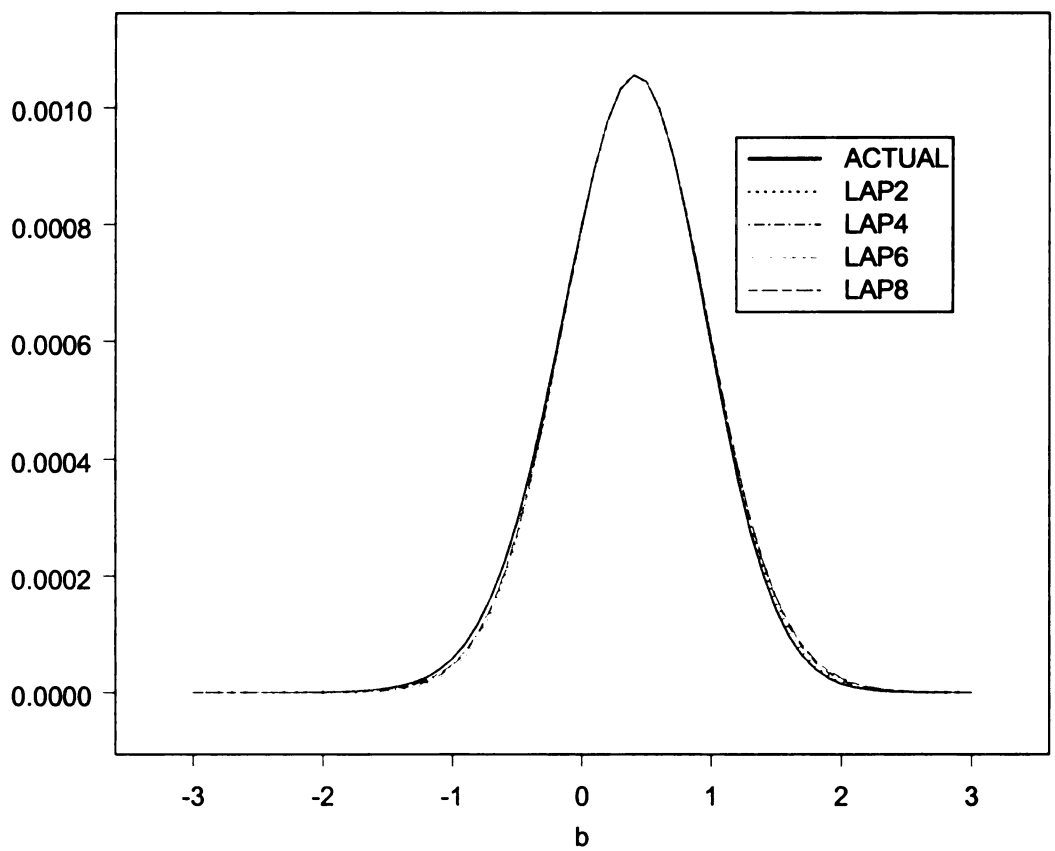
$$
\begin{aligned}
Laplace2(b) &= \exp[h(\hat{b}) - \frac{1}{2}\frac{(b-\hat{b})^2}{V}] \ , \\
Laplace4(b) &= Laplace2(b)[1+T_4] \ , \\
Laplace6(b) &= Laplace2(b)[1+T_4+T_6+\frac{1}{2}T_3^2] \ , \\
Laplace8(b) &= Laplace2(b)[1+T_4+T_6+T_8+\frac{1}{2}(T_3^2+T_4^2+2T_3T_5)] \ .
\end{aligned}
\tag{3.15}
$$

These functions are graphed along with the original integrand in Figure 1.

Using (3.12), the Laplace approximations to the integral (3.3) are given as

$$
\begin{aligned}
&Laplace2 = (2\pi V)^{1/2}\exp[h(\hat{b})] \ , \\
&Laplace4 = Laplace2[1 + E(T_4)] = Laplace2[1 - \frac{V^2}{8} n\hat{w}_i(1 - 6\hat{w}_i)] \ , \\
&Laplace6 = Laplace2[1 + E(T_4) + E(T_6) + \frac{1}{2}E(T_3^2)] \ , \\
&Laplace8 = Laplace2[1 + E(T_4) + E(T_6) + E(T_8) + \frac{1}{2}\{E(T_3^2) + E(T_4^2) + 2E(T_3T_5)\}] \ .
\end{aligned}
\tag{3.16}
$$

Figure 1: Actual integrand exp(h(b)) & Laplace Approximations



These formulas were used to compute the Laplace approximations to the integral (3.3) which are displayed in Table 3.1 along with the error from the "true" integral.

Table 3.1 - Laplace Integral Approximations

Order	Integral	Error
2	0.00145567	-0.00001765
4	0.00147026	-0.00000306
6	0.00147298	-0.00000034
8	0.00147252	-0.00000080

Gaussian Approximations

Gaussian integrand approximations. The general Gaussian integration approximation has the form

$$\int_a^b w(x)f(x)dx \approx \sum_{i=1}^{G} w_i f(x_i) \tag{3.17}$$

where $w(x)$ is a weighting function, x_i are unequally spaced abscissas (also known as quadrature points) and w_i are weights. In the Gauss-Hermite case, $w(x)=exp(-x^2)$ and $(a,b)=(-\infty,\infty)$. In this case, the abscissas x_i are obtained as the zeros of the G-th order Hermite polynomial (See Scheid, 1968; Davis and Rabinowitz, 1984). Thus

$$H_G(x)=(-1)^G e^{x^2}\frac{d^G}{dx^G}(e^{-x^2})=2^G x^G+\ \ldots \tag{3.18}$$

and the weights (coefficients) w_i are obtained using

$$w_i = \frac{2^{G+1}G!\sqrt{\pi}}{[H'_G(x_i)]^2}. \tag{3.19}$$

The numbers x_i and w_i are tabulated and widely available for various values of G (see e.g., Stroud and Sechrest, 1966). A G-order Gaussian formula requires perfect accuracy when $f(x)$ is one of the power (polynomial) functions $1, x, x^2, \ldots, x^{2G-1}$. This provides $2G$ conditions for determining $2G$ numbers x_i and w_i. In fact,

$$w_i = \int_a^b w(x) L_i(x) dx = \int_{-\infty}^{\infty} \exp[-x^2] L_i(x) dx, \tag{3.20}$$

where $L_i(x)$ is the Lagrange multiplier function (Lagrange coefficient polynomial)

$$L_i(x) = \frac{(x-x_1)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_G)}{(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_G)} = \frac{\prod_{j=1,\, j\neq i}^{G} (x-x_j)}{\prod_{j=1,\, j\neq i}^{G} (x_i-x_j)} . \tag{3.21}$$

The Lagrange formula is used as one way to find a polynomial approximation to a function $f(x)$ that collocates (i.e., coincides) with the functional values at given unequally spaced arguments. The Lagrange polynomial $p_{G-1}(x)$ of degree G-1 (or less) that passes through the G points $(x_1, f(x_1)), \ldots, (x_G, f(x_G))$ has the form

$$p_{G-1}(x) = \sum_{i=1}^{G} L_i(x) f(x_i) \tag{3.22}$$

where $L_i(x)$ is the Lagrange multiplier function defined in (3.21) having the properties

$$L_i(x_k) = 0 \text{ for } k \neq i, \; L_i(x_i) = 1. \tag{3.23}$$

Lagrange's formula represents the collocation polynomial for the unequally-spaced arguments $x_1,...,x_G$, that is,

$$p(x_k)=f(x_k) \text{ for } k=1,...,G. \tag{3.24}$$

When $p_{G-1}(x)$ is used to approximate a continuous function $f(x)$ that has G continuous derivatives, then

$$f(x)=p_{G-1}(x)+E_{G-1}(x) \tag{3.25}$$

and there exists a value $c=c(x)$ such that

$$E_{G-1}(x)=(x-x_1)...(x-x_G)\frac{f^{(G)}(c)}{G!}. \tag{3.26}$$

Thus, the Gauss-Hermite integral approximation given G abscissas and weights can be expressed, using Lagrange multiplier functions, as

$$\begin{aligned}\int_{-\infty}^{\infty}\exp[-x^2]f(x)dx \approx \sum_{i=1}^{G} w_i f(x_i) &= \sum_{i=1}^{G}\left[\int_{-\infty}^{\infty}\exp[-x^2]L_i(x)dx\right]f(x_i) \\ &= \int_{-\infty}^{\infty}\exp[-x^2]\left[\sum_{i=1}^{G}L_i(x)f(x_i)\right]dx\ .\end{aligned} \tag{3.27}$$

Essentially, this approximates the function $f(x)$ by a Lagrange polynomial. Note that the last integral is solved exactly by the Gauss-Hermite formula because $L_i(x)$ is a G-1 or less degree polynomial and Gaussian formulas (with G abscissas) give exact integrals for a

polynomial $f(x)$ up to degree $2G$-1. The integrand in (3.3), which is proportional to the likelihood $L(\beta, \tau \mid y)$, can then be written as

$$\begin{aligned}\int_{-\infty}^{\infty} \exp(h(b))db &= \int_{-\infty}^{\infty} \exp\{\sum_{i=1}^{n} [y_i(\beta+b)+\log\left(\frac{\exp[-(\beta+b)]}{1+\exp[-(\beta+b)]}\right)]-\frac{b^2}{2\tau}\}db \\ &= \int_{-\infty}^{\infty} \exp\{4(b-1)+10\ \log\left(\frac{\exp[-(b-1)]}{1+\exp[-(b-1)]}\right)-\frac{b^2}{2}\}db\ .\end{aligned} \tag{3.28}$$

Letting $x = b/\sqrt{2}$, the last integral can be written as $\sqrt{2}\int_{-\infty}^{\infty} f(x)e^{-x^2}dx$ where

$$f(x)=\exp\left\{4(\sqrt{2}x-1)+10\ \log\left(\frac{\exp(1-\sqrt{2}x)}{1+\exp(1-\sqrt{2}x)}\right)\right\}. \tag{3.29}$$

Given G quadrature points, we can use (3.27) to approximate the last integrand as

$$\int_{-\infty}^{\infty} \exp[-x^2]f(x)dx \approx \int_{-\infty}^{\infty} \exp[-x^2][\sum_{i=1}^{G} L_i(x)f(x_i)]dx. \tag{3.30}$$

The integrand on the right (the Lagrange polynomial times the weight $exp(-x^2)$) can be considered as the Gauss-Hermite integrand approximation to the integrand on the left. The Gauss-Hermite formula gives an exact value to the integral on the right. The integrand on the right was computed for number of quadrature points G ranging from 2 to 30 and plotted against the argument. The plots, along with the one for the actual integrand (the integrand on the left), are displayed in Figures 2--5. Note that as the number of

quadrature points increases, the graph becomes practically indistinguishable from the one for the actual integrand.

Figure 2: Plot of actual integrand & GH approximations

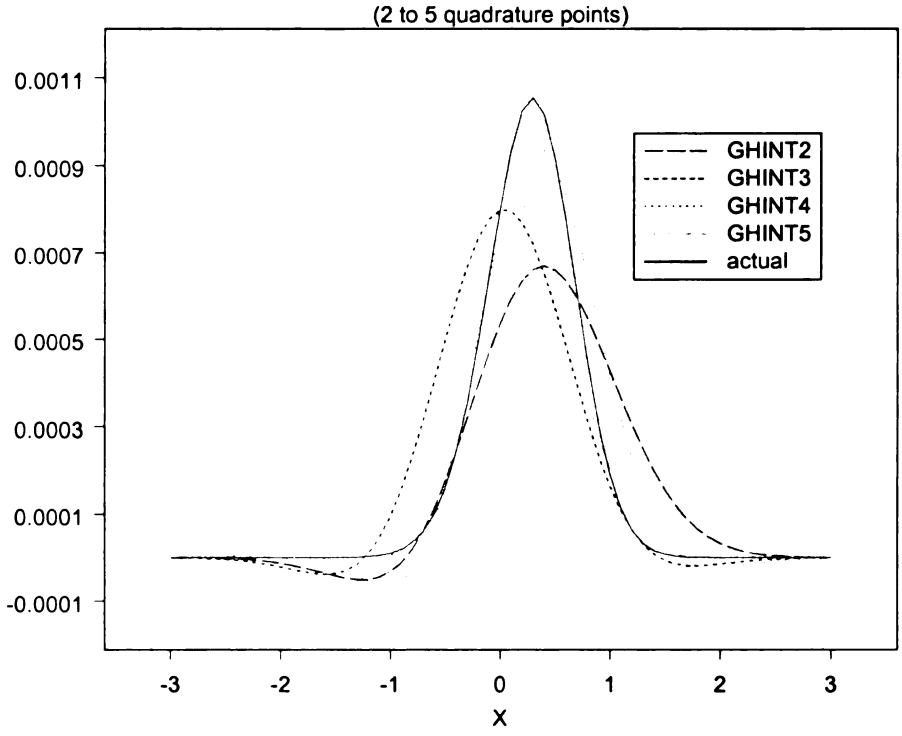


Figure 3: Actual integrand & GH approximations

(6 to 10 quadrature points)

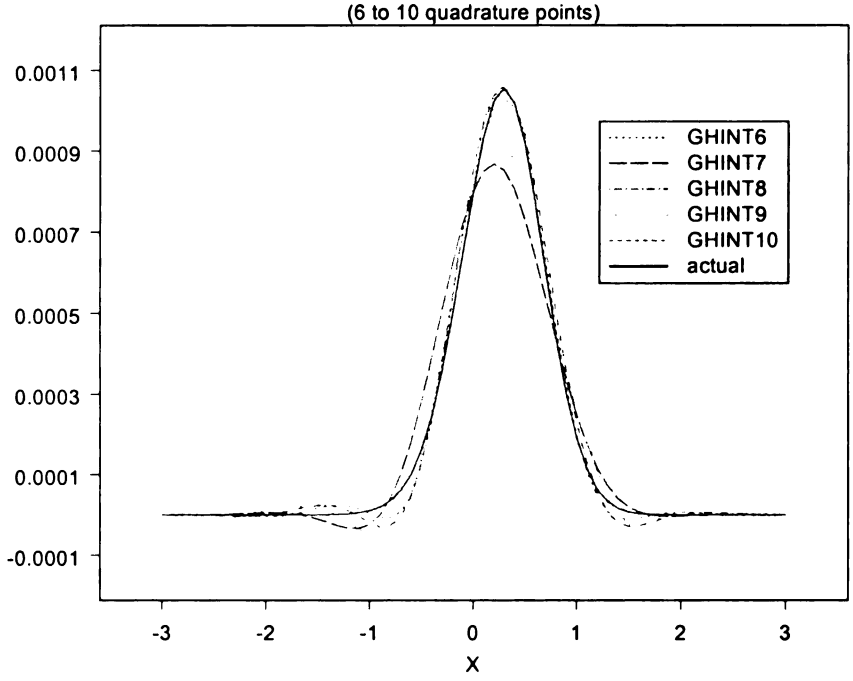


Figure 4: Actual integrand & GH approximations

(16 to 20 quadrature points)

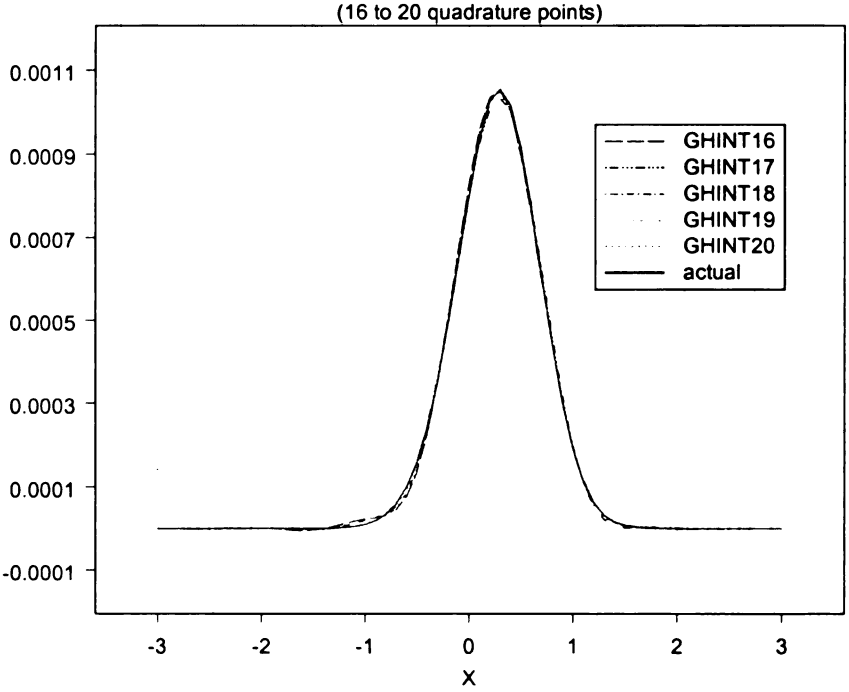
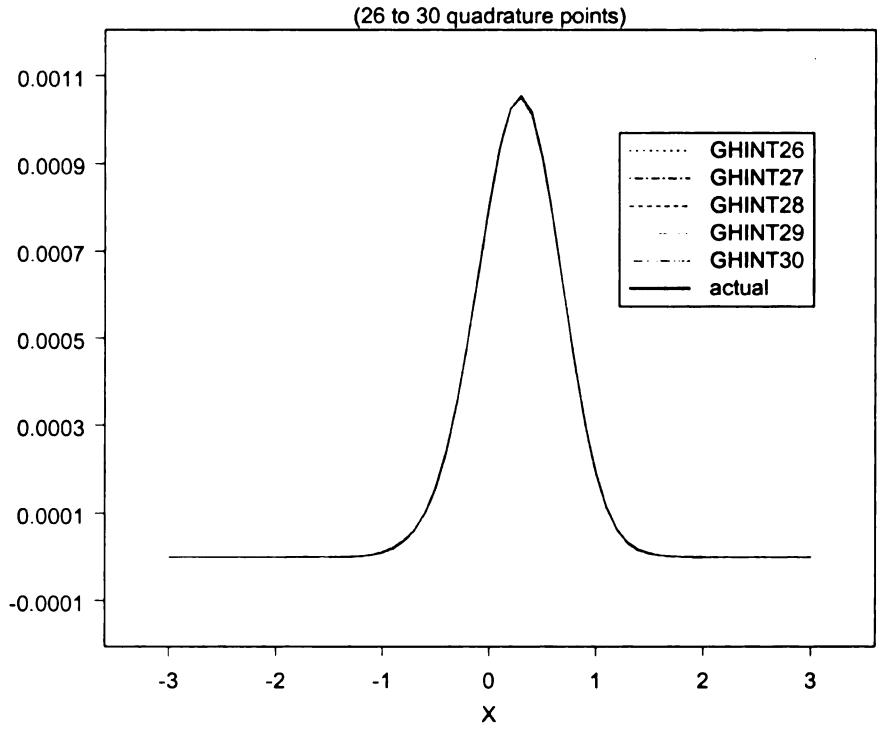


Figure 5: Actual integrand & GH approximations



For the *adaptive* Gauss-Hermite, we first expand $h(b)$ in a second-order Taylor expansion around its maximizer $\hat{b}$, i.e.,

$$h(b) \approx h(\hat{b}) + \frac{1}{2}(b-\hat{b})' h''(\hat{b})(b-\hat{b}). \tag{3.31}$$

By substituting this in (3.3), we note that, up to a multiplicative constant, b can be considered as $N(\hat{b}, -[h''(\hat{b})]^{-1}) = N(.41717, -[h''(.41717)]^{-1})$. Let

$$b = \mu_{\hat{b}} + \sigma_{\hat{b}} z = \hat{b} + [-h''(\hat{b})]^{-1/2} z \Rightarrow db = [-h''(\hat{b})]^{-1/2} dz. \tag{3.32}$$

The integral in (3.3) can then be rewritten as

$$\begin{aligned}&\int_{-\infty}^{\infty}[-h''(\hat{b})]^{-1/2}\exp\{h(\hat{b}+[-h''(\hat{b})]^{-1/2}z)+\frac{z^2}{2}\}\exp(-\frac{z^2}{2})dz\\&=\sqrt{2}[-h''(\hat{b})]^{-1/2}\int_{-\infty}^{\infty}\exp\{h(\hat{b}+\sqrt{2}[-h''(\hat{b})]^{-1/2}x)+x^2\}\exp[-x^2]dx\end{aligned} \tag{3.33}$$

which is approximated, using the G-point Gauss-Hermite formula, by

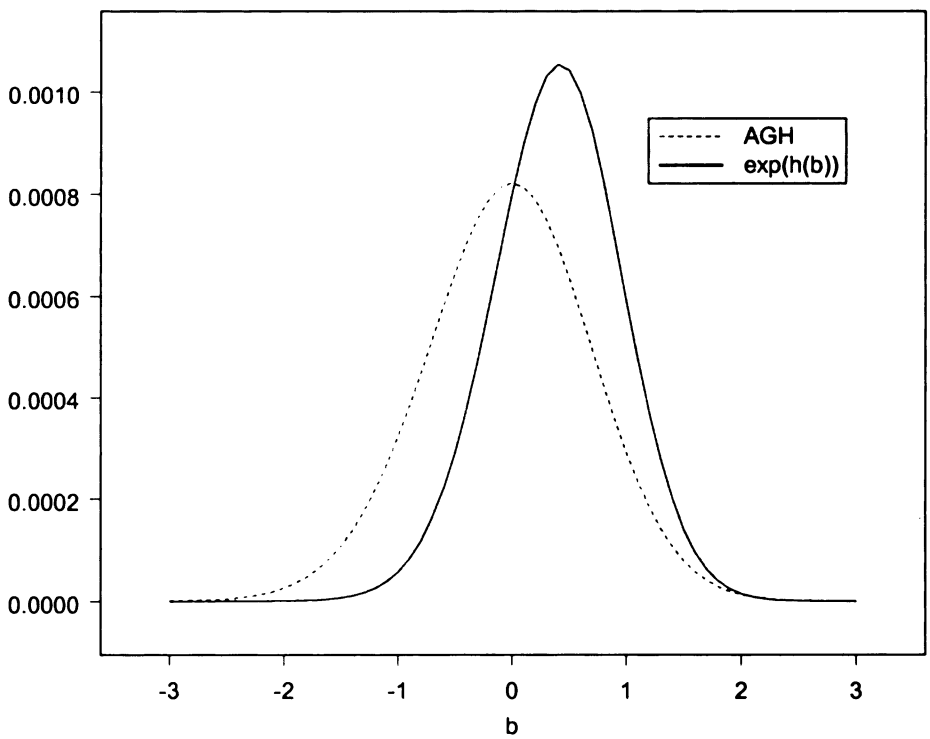
$$\sqrt{2}[-h''(\hat{b})]^{-1/2}\sum_{k=1}^{G}w_k\exp\{h(\hat{b}+\sqrt{2}[-h''(\hat{b})]^{-1/2}x_k)+x_k^2\}. \tag{3.34}$$

This is the adaptive Gauss-Hermite approximation to the integral (3.3). Note that for G=1, (3.18) and (3.19) give x=0 as the only abscissa and $w_1 = \sqrt{\pi}$ as the corresponding weight. Thus, for G=1, (3.34) reduces to $\sqrt{2\pi}[-h''(\hat{b})]^{-1/2}\exp\{h(\hat{b})\}$ which is identical to Laplace2 given in (3.16).

The integrand in (3.33), which I call the AGH integrand, is plotted along with the true integrand (the integrand in (3.3)) in Figure 6. The AGH integrand is the transformed (standardized) version of the true integrand in the sense of standardizing a normal random variable. Note that, unlike the true integrand, the AGH integrand is centered around zero. Like the non-adaptive Gauss-Hermite case, (3.27) can be used to obtain approximations

to the integrand in (3.33) using the Lagrange multiplier function where, now, $f(x)$ is the integrand in (3.33) without the weight function $exp(-x^2)$. This integrand approximation to the AGH integrand using the Lagrange multiplier function has been computed for

Figure 6: Graphs of true integrand exp(h(b)) & AGH integrand



numbers of quadrature points varying from 2 to 30 but plotted for up to 10 quadrature points. Figures 7 and 8 display the plots. As can be seen from the plots, the graphs converge fast (faster than the non-adaptive GH) to that of the AGH integrand.

Figure 7: AGH integrand & AGH2-AGH5 approximations

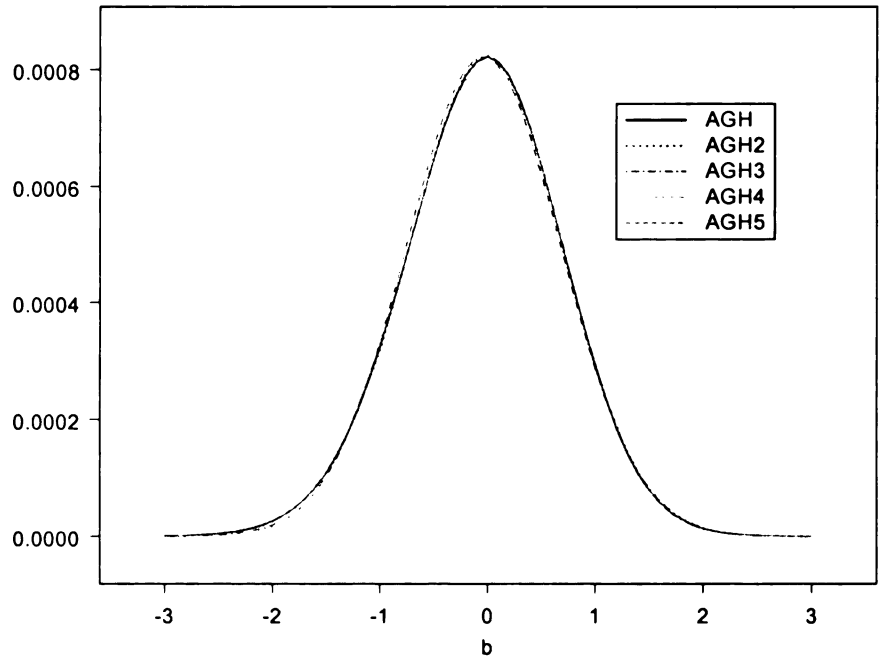
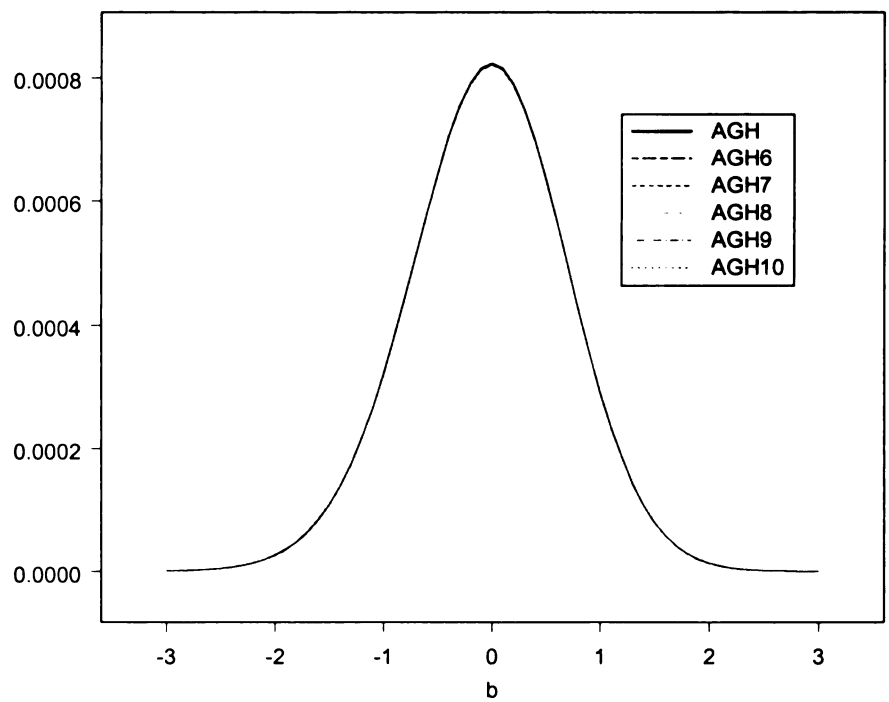
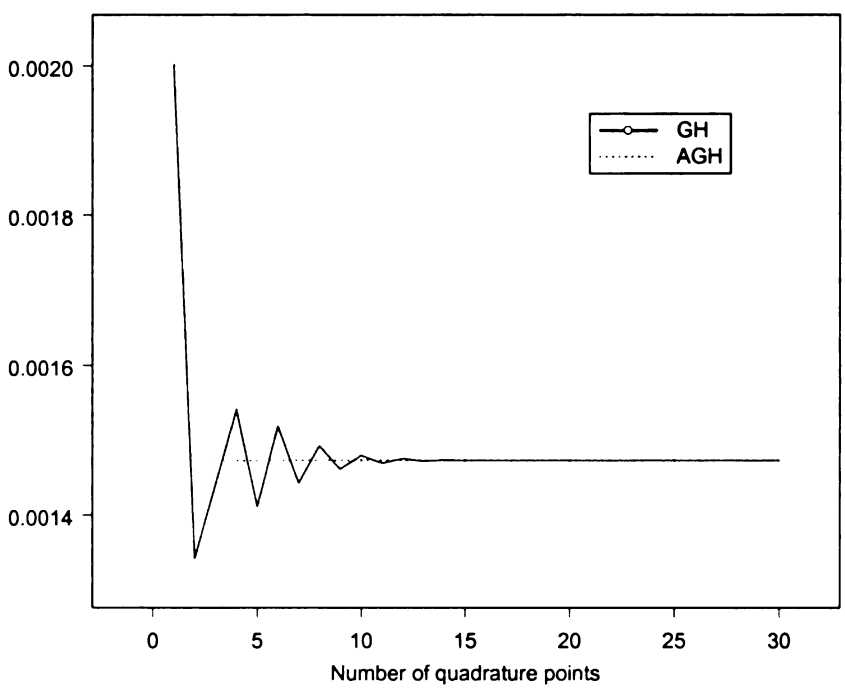


Figure 8: AGH integrand & AGH6-AGH10 approximations



Gaussian integral approximations. The Gauss-Hermite integral approximations to the integral in (3.3), both non-adaptive and adaptive, were computed for numbers of quadrature points varying from 1 to 30 and tabulated for 1 to 27 quadrature points. Equation (3.17), with f defined as in (3.29), was used to compute the (non-adaptive) Gauss-Hermite integral approximations and formula (3.34) was used to compute the adaptive counterparts. The results, including the error from the "true" integral value, are given in Table 3.2. Note that the adaptive Gauss-Hermite integral approximation with one quadrature point is the same as the Laplace2 integral approximation. The Gaussian integral approximation values are also plotted in Figure 9, this time against the

Figure 9 : Gaussian integral approximations vs no. of quadrature points



number of quadrature points, in order to compare the convergence behaviors of the two approaches. The figure displays that the adaptive GH converges to the true value faster (and without zigzagging).

Table 3.2 - Gaussian Integral Approximations

G	GH Integral	Error	AGH Integral	Error
1	0.00200186	0.00052854	0.00145567	-0.00001765
2	0.00134210	-0.00013122	0.00146071	-0.00001261
3	0.00144044	-0.00003288	0.00147185	-0.00000147
4	0.00154149	0.00006817	0.00147303	-0.00000029
5	0.00141145	-0.00006187	0.00147323	-0.00000009
6	0.00151908	0.00004577	0.00147331	-0.00000001
7	0.00144266	-0.00003066	0.00147331	-0.00000001
8	0.00149265	0.00001933	0.00147332	-0.00000000
9	0.00146166	-0.00001166	0.00147332	-0.00000000
10	0.00148010	0.00000678	0.00147332	-0.00000000
11	0.00146952	-0.00000380	0.00147332	-0.00000000
12	0.00147537	0.00000205	0.00147332	0.00000000
13	0.00147226	-0.00000106	0.00147332	0.00000000
14	0.00147383	0.00000051	0.00147332	0.00000000
15	0.00147310	-0.00000022	0.00147332	0.00000000
16	0.00147339	0.00000008	0.00147332	0.00000000
17	0.00147331	-0.00000001	0.00147332	0.00000000
18	0.00147330	-0.00000002	0.00147332	0.00000000
19	0.00147335	0.00000003	0.00147332	0.00000000
20	0.00147329	-0.00000003	0.00147332	0.00000000
21	0.00147334	0.00000002	0.00147332	0.00000000
22	0.00147330	-0.00000002	0.00147332	0.00000000
23	0.00147333	0.00000001	0.00147332	0.00000000
24	0.00147331	-0.00000001	0.00147332	0.00000000
25	0.00147332	0.00000001	0.00147332	0.00000000
26	0.00147331	-0.00000000	0.00147332	0.00000000
27	0.00147332	0.00000000	0.00147332	0.00000000

Chapter 4

METHOD

Introduction

In this chapter, several methods for approximating the marginal likelihood are discussed and compared to each other analytically. The methods compared here are those that involve centering around the conditional mode, namely standard Laplace (Laplace2 or PQL), Laplace6, and the adaptive Gaussian quadrature methods. Since the non-adaptive Gauss-Hermite is centered around zero (rather than the conditional mode) and was shown to be inferior to the adaptive in terms of its efficiency (see Pinheiro and Bates, 1995), it will not be compared here. However, the formula used by it, as well as the other methods, to approximate the integral will be given. First, the model being estimated will be formulated. Next, formulas will be given for all the methods considered here to show how they approximate the integral to obtain approximate likelihood.

The Model

Let Y_{ij} be the response variable representing the outcome of a level-1 unit (e.g., subject) i in level-2 unit (cluster) j, $i=1,...,n_j$; $j=1,...,J$. Let b_j be a random cluster effect, and let $\mu_{ij}=E(Y_{ij}|b_j)$. Assume that the distribution of $Y_{ij}|b_j$ is a member of the exponential family, i.e.,

$$f_{y_{ij}|b_j}(y_{ij}|b_j)=\exp\{[y_{ij}\eta_{ij}-\delta(\eta_{ij})]/\alpha(\phi)+\gamma(y_{ij},\phi)\} \qquad \textbf{(4.1)}$$

for some functions α, δ and γ, where η_{ij} is the canonical parameter and φ is called the dispersion parameter (see McCullagh and Nelder, 1989, p.28). The outcome Y_{ij} is related to the fixed and random effects through its conditional mean via the canonical link function η_{ij} of μ_{ij}. Thus,

$$\eta_{ij} = x_{ij}'\beta + z_{ij}'b_j \ , \tag{4.2}$$

where x_{ij} is the $p\times 1$ covariate vector and z_{ij} is the $r\times 1$ design vector for the r random effects, β is the $p\times 1$ vector of unknown fixed regression coefficients, b_j is the $r\times 1$ vector of random effects assumed to be distributed $N_r(0,D)$. In matrix form,

$$\eta_j = X_j\beta + Z_j b_j \ . \tag{4.3}$$

Let $y_j = (y_{1j}, y_{2j}, \ldots, y_{n_j j})'$ be the vector of responses for the n_j level-1 units nested within level-2 unit j. Given the conditional distribution of y_j given b_j (and β), the marginal likelihood is given by

$$L(\beta, D;\ y) = \prod_{j=1}^{J} f_{y_j}(y_j) \tag{4.4}$$

where

$$\begin{aligned} f_{y_j}(y_j) &= \int_{-\infty}^{\infty} f_{y_j|b_j}(y_j|b_j)g(b_j)db_j \\ &= (2\pi)^{-r/2}|D|^{-1/2}\int_{-\infty}^{\infty} f_{y_j|b_j}(y_j|b_j)\exp[-\frac{1}{2}b_j'D^{-1}b_j]db_j \end{aligned} \ . \tag{4.5}$$

For a binary outcome and logit link, the conditional distribution of y_j given b_j is given by

$$f_{y_j|b_j}(y_j|b_j)=\prod_{i=1}^{n_j}\mu_{ij}^{y_{ij}}(1-\mu_{ij})^{1-y_{ij}}\ , \tag{4.6}$$

where the conditional mean $\mu_{ij} = E(y_{ij}|b_j)$ is related to the fixed and random effects

via the logit link $\eta_{ij} = \log(\frac{\mu_{ij}}{1-\mu_{ij}})$, whence

$$\begin{aligned}\mu_{ij}&=\frac{1}{1+\exp(-\eta_{ij})}=\frac{1}{1+\exp[-(x_{ij}'\beta+z_{ij}'b_j)]}\ ,\\ \frac{d\mu_{ij}}{d\eta_{ij}}&=\frac{\exp(-\eta_{ij})}{[1+\exp(-\eta_{ij})]^2}=\mu_{ij}(1-\mu_{ij})\ .\end{aligned} \tag{4.7}$$

The integrand in (4.5) can be written as $exp[h(b_j)]$ where

$$\begin{aligned}h(b_j)&=\log f_{y_j|b_j}(y_j|b_j)-\frac{1}{2}b_j'D^{-1}b_j\\ &=\sum_{i=1}^{n_j}[y_{ij}\log(\mu_{ij})+(1-y_{ij})\log(1-\mu_{ij})]-\frac{1}{2}b_j'D^{-1}b_j\ .\\ &=\sum_{i=1}^{n_j}[y_{ij}\eta_{ij}+\log(1-\mu_{ij})]-\frac{1}{2}b_j'D^{-1}b_j\end{aligned} \tag{4.8}$$

Let $\hat{b}_j = \hat{b}_j(\beta, D, y_j)$ be the value of b_j that maximizes $h(b_j)$. The integral in (4.5)

can then be written as

$$\int_{-\infty}^{\infty}\exp[h(b_j)]db_j=\exp[h(\hat{b}_j)]\int_{-\infty}^{\infty}\exp[-\frac{1}{2}(b_j-\hat{b}_j)'V_j^{-1}(b_j-\hat{b}_j)]\exp[S]db_j$$
$$=(2\pi)^{r/2}|V_j|^{1/2}\exp[h(\hat{b}_j)]E(\exp[S]) \qquad (4.9)$$

where $h'(\hat{b}_j)$ vanishes because $\hat{b}_j$ is the maximizer and

$$V_j=-[h''(\hat{b}_j)]^{-1}=-\left[\frac{\partial^2 h(b_j)}{\partial b_j \partial b_j'}|_{b_j=\hat{b}_j}\right]^{-1};$$
$$S=\sum_{k=3}^{\infty}T_{kj}\ ,\quad T_{kj}=\frac{1}{k!}[\overset{k-1}{\otimes}(b_j-\hat{b}_j)']h^{(k)}(\hat{b}_j)(b_j-\hat{b}_j) \qquad (4.10)$$

Here, $\overset{k}{\otimes}x=x\otimes x\otimes\cdots\otimes x$ is used to mean the Kronecker product of k x's (see Raudenbush, Yang and Yosef, 2000).

The approaches used by the various methods compared here to approximate the above integral are given subsequently. Prior to that, let's derive the first two derivatives of h.

$$h'(b_j)=\frac{\partial h(b_j)}{\partial b_j}=\sum_{i=1}^{n_j}\left[y_{ij}\frac{\partial\eta_{ij}}{\partial b_j}+\frac{1}{1-\mu_{ij}}\left(-\frac{\partial\mu_{ij}}{\partial b_j}\right)\right]-D^{-1}b_j$$
$$=\sum_{i=1}^{n_j}\left(y_{ij}z_{ij}-\frac{1}{1-\mu_{ij}}\frac{d\mu_{ij}}{d\eta_{ij}}\frac{\partial\eta_{ij}}{\partial b_j}\right)-D^{-1}b_j \qquad (4.11)$$
$$=\sum_{i=1}^{n_j}(y_{ij}z_{ij}-\mu_{ij}z_{ij})-D^{-1}b_j=\sum_{i=1}^{n_j}(y_{ij}-\mu_{ij})z_{ij}-D^{-1}b_j$$

and

$$h''(b_j) = \frac{\partial^2 h(b_j)}{\partial b_j \partial b_j'} = -\sum_{i=1}^{n_j} \mu_{ij}(1-\mu_{ij}) z_{ij} z_{ij}' - D^{-1} \\ = -(Z_j' W_j Z_j + D^{-1}) \tag{4.12}$$

where W_j is an $n_j \times n_j$ diagonal matrix with $w_{ij} = \mu_{ij}(1 - \mu_{ij})$ on the diagonal.

Laplace Approximations

Standard Laplace (*Laplace2*)

Laplace2 essentially expands $h(b_j)$ in a second order Taylor series around $\hat{b}_j$

and then uses Normal theory to complete the integration, i.e., in the formula (4.9) it approximates $E(e^S) \approx 1$. Hence, the L approximation for the integral (4.5) is

$(2\pi)^{r/2} |V_j|^{1/2} \exp[h(\hat{b}_j)]$ and the marginal likelihood is approximated as

$$Laplace2 \; : \; L \approx |D|^{-J/2} \prod_{j=1}^{J} |V_j|^{1/2} \exp[h(\hat{b}_j)] \tag{4.13}$$

and the log-likelihood as

$$\log(L) \approx -\frac{J}{2}\log|D| + \frac{1}{2}\sum_{j=1}^{J} \log|V_j| + \sum_{j=1}^{J} h(\hat{b}_j) \; . \tag{4.14}$$

Sixth-order Laplace (*Laplace6*)

The sixth-order Laplace approximation (Yang, 1998; Raudenbush et al., 2000) is based on the approximation

$$E(e^{S}) \approx E(1 + S + \frac{1}{2}S^2) \approx 1 + E(T_4) + E(T_6) + \frac{1}{2}E(T_3^2) \qquad \textbf{(4.15)}$$

noting that, based on Normal theory, $E(T_k) = 0$ for odd k. Thus, the sixth-order Laplace approximation to the likelihood becomes (see Raudenbush et al., 2000)

$$Laplace6 \;:\; L \;\approx\; |D|^{-J/2} \prod_{j=1}^{J} |V_j|^{1/2} \exp[h(\hat{b}_j)][1 + E(T_{4j}) + E(T_{6j}) + \frac{1}{2}E(T_{3j}^2)] \qquad \textbf{(4.16)}$$

whence the log-likelihood becomes

$$\log(L) \;\approx\; \frac{-J}{2}\log|D| + \frac{1}{2}\sum_{j=1}^{J} \log|V_j| + \sum_{j=1}^{J} h(\hat{b}_j) + \sum_{j=1}^{J} \log[1 + E(T_{4j}) + E(T_{6j}) + \frac{1}{2}E(T_{3j}^2)] \;. \qquad \textbf{(4.17)}$$

Gaussian Quadrature Approximations

Non-adaptive Gauss-Hermite Quadratures

The Gauss-Hermite integration approximation formula numerically approximates an integral whose integrand is a product of a function $f(x)$ and $\exp(-x'x)$ and whose limits of integration are -∞ and ∞. In the univariate case, the *G*-point Gauss-Hermite integral approximation formula is given by

$$\int_{-\infty}^{\infty} f(x)\exp(-x^2)dx \approx \sum_{g=1}^{G} w_g f(x_g) \qquad \textbf{(4.18)}$$

where $x_g, w_g, g = 1,\ldots,G$ are, respectively, the G Gaussian quadrature points (abscissas) and the corresponding weights. The abscissas and the weights are determined in such a way that the formula (4.18) is exact for polynomials up to degree $2G$-1. The abscissas are the zeros of the Gth degree Hermite polynomial and the weights are functions of the Gth degree Hermite polynomial at x_g, $H_G(x_g)$ (See Chapter 3). Both the abscissas x_g and the weights w_g are tabulated for varying values of G and tables are widely available (e.g., Stroud and Sechrest, 1966). The extension to the multivariate x case follows naturally.

To make the integral in (4.5) amenable to the Gauss-Hermite approximation, we need a transformation. Let $u_j = T^{-1}b_j / \sqrt{2}$, where $TT' = D$ is the Cholesky decomposition of D. This implies $b_j = \sqrt{2}\, Tu_j$, and

$$\left|\frac{db_j}{du_j'}\right| = 2^{r/2}|T| = 2^{r/2}|D|^{1/2}. \qquad \textbf{(4.19)}$$

The integral in (4.5) can then be written as

$$f_{y_j}(y_j) = \pi^{-r/2} \int_{-\infty}^{\infty} \prod_{i=1}^{n_j} \mu_{ij}^{y_{ij}} (1-\mu_{ij})^{1-y_{ij}} \exp(-u_j' u_j) du_j$$
$$= \pi^{-r/2} \int_{-\infty}^{\infty} \exp\left\{ \sum_{i=1}^{n_j} [y_{ij}\eta_{ij} + \log(1-\mu_{ij})] \right\} \exp(-u_j' u_j) du_j \qquad \textbf{(4.20)}$$

where, now,

$$\mu_{ij} = \frac{1}{1+\exp(-\eta_{ij})} = \frac{1}{1+\exp[-(x_{ij}'\beta + \sqrt{2}\, z_{ij}' T u_j)]} . \qquad \textbf{(4.21)}$$

The Gauss-Hermite approximation to the integral is given by

$$\sum_{g_1=1}^{G} \sum_{g_2=1}^{G} \cdots \sum_{g_r=1}^{G} (w_{g_1} w_{g_2} \cdots w_{g_r}) \prod_{i=1}^{n_j} [\mu_{ij}(u_{jg_1}, \ldots, u_{jg_r})]^{y_{ij}} [1-\mu_{ij}(u_{jg_1}, \ldots, u_{jg_r})]^{1-y_{ij}} \qquad \textbf{(4.22)}$$

where

$$\mu_{ij}(u_{jg_1}, \ldots, u_{jg_r}) = \frac{1}{1+\exp\left[-\left(x_{ij}'\beta + \sqrt{2}\, z_{ij}' T \begin{pmatrix} u_{jg_1} \\ \cdots \\ u_{jg_r} \end{pmatrix}\right)\right]} . \qquad \textbf{(4.23)}$$

Adaptive Gauss-Hermite Quadratures

A second-order Taylor expansion of $h(b_j)$ around its maximizer $\hat{b}_j$ gives

$$h(b_j) \approx h(\hat{b}_j) + \frac{1}{2}(b_j - \hat{b}_j)' h''(\hat{b}_j)(b_j - \hat{b}_j) \ . \qquad \textbf{(4.24)}$$

By substituting the second-order Taylor expansion for $h(b_j)$ in the integral on the left hand side of (4.9), we note that up to a multiplicative constant, b_j can be thought of as distributed $N_r(\hat{b}_j, -[h''(\hat{b}_j)]^{-1})$. Let $z \sim N_r(0, I)$ and $b_j = \mu_{b_j} + \Sigma_{b_j} z = \hat{b}_j + [-h''(\hat{b}_j)]^{-1/2} z$. Then,

$$\left| \frac{\partial b_j}{\partial z'} \right| = [-h''(b_j)]^{-1/2} \ . \qquad \textbf{(4.25)}$$

Following Pinheiro and Bates (1995), the left-hand side integral of (4.9) can be written as

$$\begin{aligned} &\int_{-\infty}^{\infty} [-h''(\hat{b}_j)]^{-1/2} \exp\left\{ h(\hat{b}_j + [-h''(\hat{b}_j)]^{-1/2} z) + \frac{z'z}{2} \right\} \exp[-\frac{z'z}{2}] dz \\ &= \sqrt{2}[-h''(\hat{b}_j)]^{-1/2} \int_{-\infty}^{\infty} \exp\left\{ h(\hat{b}_j + \sqrt{2}[-h''(\hat{b}_j)]^{-1/2} u_j) + u_j' u_j \right\} \exp[-u_j' u_j] du_j \end{aligned} \qquad \textbf{(4.26)}$$

where $u_j = z / \sqrt{2}$. The last integral is approximated, using the G-point Gauss-Hermite formula, by

$$\sqrt{2}[-h''(\hat{b}_j)]^{-1/2}\sum_{g_1=1}^{G}\sum_{g_2=1}^{G}\cdots\sum_{g_r=1}^{G}(w_{g_1}w_{g_2}\cdots w_{g_r})\times$$
$$\exp\left\{h\left(\hat{b}_j+\sqrt{2}[-h''(\hat{b}_j)]^{-1/2}\begin{pmatrix}u_{jg_1}\\ \cdots\\ u_{jg_r}\end{pmatrix}\right)+(u_{jg_1},\ldots,u_{jg_r})\begin{pmatrix}u_{jg_1}\\ \cdots\\ u_{jg_r}\end{pmatrix}\right\} \quad \textbf{(4.27)}$$

The Single Random Effect Case

When r=1, i.e., the random effects term is univariate, the formulas for the four methods simplify as follows. The Laplace2 approximation to the integral reduces to

$$(2\pi V_j)^{1/2}\exp[h(\hat{b}_j)] = [2\pi/-h''(\hat{b}_j)]^{1/2}\exp[h(\hat{b}_j)] \quad \textbf{(4.28)}$$

and the corresponding sixth order Laplace approximation becomes

$$\begin{aligned}&(2\pi V_j)^{1/2}\exp[h(\hat{b}_j)]\times[1+E(T_{4j})+E(T_{6j})+\frac{1}{2}E(T_{3j}^2)]\\ &= [2\pi/-h''(\hat{b}_j)]^{1/2}\exp[h(\hat{b}_j)]\times[1+E(T_{4j})+E(T_{6j})+\frac{1}{2}E(T_{3j}^2)]\end{aligned} \quad \textbf{(4.29)}$$

where

$$T_{kj}=\frac{1}{k!}h^{(k)}(\hat{b}_j)(b_j-\hat{b}_j)^k \quad \textbf{(4.30)}$$

is the kth Taylor expansion term and expectations are taken over $N(0, -[h''(\hat{b}_j)]^{-1})$.

The Gauss-Hermite approximation to the integral reduces to

$$\sum_{g=1}^{G} w_g \prod_{i=1}^{n_j} [\mu_{ij}(u_{jg})]^{y_{ij}} [1-\mu_{ij}(u_{jg})]^{1-y_{ij}} = \sum_{g=1}^{G} w_g \exp\left\{\sum_{i=1}^{n_j} [y_{ij}\eta_{ij}(u_{jg}) + \log(1-\mu_{ij}(u_{jg}))]\right\} \quad \textbf{(4.31)}$$

where

$$\mu_{ij}(u_{jg}) = [1+\exp[-\eta_{ij}(u_{jg})]]^{-1} = [1+\exp[-(x_{ij}'\beta + \sqrt{2D}\ u_{jg})]^{-1} \quad \textbf{(4.32)}$$

and the adaptive Gaussian approximation is given by

$$\sqrt{2}[-h''(\hat{b}_j)]^{-1/2} \sum_{g=1}^{G} w_g \exp\left\{h(\hat{b}_j) + \sqrt{2}[-h''(\hat{b}_j)]^{-1/2} u_{jg} + u_{jg}^2\right\}. \quad \textbf{(4.33)}$$

Asymptotic Behavior of the Methods

For simplicity, let $n_j = n$ for all j. For all practical purposes, the four methods can be considered as approximations to the integral on the left hand side of (4.9), which can be written as

$$\int_{-\infty}^{\infty} \exp[h(b_j)]db_j = \int_{-\infty}^{\infty} \exp[n\ l(b_j)]db_j = \int_{-\infty}^{\infty} g(b_j)db_j \quad \textbf{(4.34)}$$

where

$$l(b_j) = \frac{1}{n}\left[\sum_i y_{ij}\eta_{ij} + \sum_i \log(1-\mu_{ij}) - \frac{b_j^2}{2D}\right]. \quad \textbf{(4.35)}$$

That means the approximation to the integral is done for each cluster j. So, the errors are derived for a single cluster.

Tierney and Kadane (1986) have shown that the error of the standard Laplace approximation method (called Laplace2 or PQL here) is of order $O(n^{-1})$. This can also be easily shown using the definitions of chapter 3 and formulas given on p.147 of Raudenbush et al. (2000).

To derive the order for the error for Laplace6, we note that since the terms in S of (4.10) diminish as a function of the cluster size n (Raudenbush et al., 2000), the error is of the same order as of the terms not included in Laplace6 but included in the next higher-order Laplace approximation. In other words, the error of Laplace6 is of the same order as the error of the terms that result from the difference of Laplace8 and Laplace6 (see (3.12) in Chapter 3 for these terms and definition of Laplace8). From chapter 3, we see that for a given cluster j, the difference between Laplace8 and Laplace6 is

$$Laplace8 - Laplace6 = E(T_{8j}) + \frac{1}{2}E(T_{4j}^2) + E(T_{3j}T_{5j}) \qquad \textbf{(4.36)}$$

where T_{kj} is as defined in (4.30). From Raudenbush et al (2000, p.147), for the univariate case, we have

$$h^{(k)}(\hat{b}_j) = -\sum_{i=1}^{n} \hat{m}_{ij}^{(k)} = -\sum_{i=1}^{n} \frac{\partial^{k-1}\mu_{ij}}{\partial\eta_{ij}^{k-1}}(\hat{b}_j), \text{ for } k \geq 3. \qquad \textbf{(4.37)}$$

The derivatives are given in (4.2) of Raudenbush et al. (2000) for k=3 to 6. Thus,

$$
\begin{aligned}
E(T_{4j}^2) &= E\left[\frac{h^{(4)}(\hat{b}_j)}{4!}(b_j-\hat{b}_j)^4\right]^2 = [-\sum_{i=1}^{n}\hat{m}_{ij}^{(4)}/4!]^2 E(b_j-\hat{b}_j)^8 \\
&= (\frac{n}{4!}\sum_{i=1}^{n}\hat{w}_{ij}(1-6\hat{w}_{ij})/n)^2\mu_8 \\
&= O(n^2)7\cdot5\cdot3\sigma^8 \;=\; O(n^2)105[-h''(\hat{b}_j)]^{-4} \\
&= O(n^2)105[-n(\sum_{i=1}^{n}\hat{w}_{ij}+D^{-1})/n]^{-4} = O(n^2)O(n^{-4}) = O(n^{-2})
\end{aligned}
\tag{4.38}
$$

Likewise,

$$
\begin{aligned}
E(T_{3j}T_{5j}) &= E\left[\frac{h^{(3)}(\hat{b}_j)}{3!}(b_j-\hat{b}_j)^3\frac{h^{(5)}(\hat{b}_j)}{5!}(b_j-\hat{b}_j)^5\right] \\
&= \frac{105}{3!5!}h^{(3)}(\hat{b}_j)h^{(5)}(\hat{b}_j)O(n^{-4}) \\
&= \frac{105}{3!5!}[-n\sum_{i=1}^{n}\hat{w}_{ij}(1-2\hat{\mu}_{ij})/n][-n\sum_{i=1}^{n}\hat{w}_{ij}(1-2\hat{\mu}_{ij})(1-12\hat{w}_{ij})/n]O(n^{-4}) \\
&= O(n^2)O(n^{-4}) = O(n^{-2})
\end{aligned}
\tag{4.39}
$$

and

$$
\begin{aligned}
E(T_{4j}^2) &= E\left[\frac{h^{(4)}(\hat{b}_j)}{4!}(b_j-\hat{b}_j)^4\right]^2 = [-\sum_{i=1}^{n}\hat{m}_{ij}^{(4)}/4!]^2 E(b_j-\hat{b}_j)^8 \\
&= (\frac{n}{4!}\sum_{i=1}^{n}\hat{w}_{ij}(1-6\hat{w}_{ij})/n)^2\mu_8 \\
&= O(n^2)7\cdot5\cdot3\sigma^8 \;=\; O(n^2)105[-h''(\hat{b}_j)]^{-4} \\
&= O(n^2)105[-n(\sum_{i=1}^{n}\hat{w}_{ij}+D^{-1})/n]^{-4} = O(n^2)O(n^{-4}) = O(n^{-2})
\end{aligned}
\tag{4.38}
$$

Likewise,

$$
\begin{aligned}
E(T_{3j}T_{5j}) &= E\left[\frac{h^{(3)}(\hat{b}_j)}{3!}(b_j-\hat{b}_j)^3\frac{h^{(5)}(\hat{b}_j)}{5!}(b_j-\hat{b}_j)^5\right] \\
&= \frac{105}{3!5!}h^{(3)}(\hat{b}_j)h^{(5)}(\hat{b}_j)O(n^{-4}) \\
&= \frac{105}{3!5!}[-n\sum_{i=1}^{n}\hat{w}_{ij}(1-2\hat{\mu}_{ij})/n][-n\sum_{i=1}^{n}\hat{w}_{ij}(1-2\hat{\mu}_{ij})(1-12\hat{w}_{ij})/n]O(n^{-4}) \\
&= O(n^2)O(n^{-4}) = O(n^{-2})
\end{aligned}
\tag{4.39}
$$

and

$$
\begin{aligned}
E(T_{4j}^2) &= E\left[\frac{h^{(4)}(\hat{b}_j)}{4!}(b_j-\hat{b}_j)^4\right]^2 = [-\sum_{i=1}^{n}\hat{m}_{ij}^{(4)}/4!]^2 E(b_j-\hat{b}_j)^8 \\
&= (\frac{n}{4!}\sum_{i=1}^{n}\hat{w}_{ij}(1-6\hat{w}_{ij})/n)^2\mu_8 \\
&= O(n^2)7\cdot5\cdot3\sigma^8 \;=\; O(n^2)105[-h''(\hat{b}_j)]^{-4} \\
&= O(n^2)105[-n(\sum_{i=1}^{n}\hat{w}_{ij}+D^{-1})/n]^{-4} = O(n^2)O(n^{-4}) = O(n^{-2})
\end{aligned}
\tag{4.38}
$$

Likewise,

$$
\begin{aligned}
E(T_{3j}T_{5j}) &= E\left[\frac{h^{(3)}(\hat{b}_j)}{3!}(b_j-\hat{b}_j)^3\frac{h^{(5)}(\hat{b}_j)}{5!}(b_j-\hat{b}_j)^5\right] \\
&= \frac{105}{3!5!}h^{(3)}(\hat{b}_j)h^{(5)}(\hat{b}_j)O(n^{-4}) \\
&= \frac{105}{3!5!}[-n\sum_{i=1}^{n}\hat{w}_{ij}(1-2\hat{\mu}_{ij})/n][-n\sum_{i=1}^{n}\hat{w}_{ij}(1-2\hat{\mu}_{ij})(1-12\hat{w}_{ij})/n]O(n^{-4}) \\
&= O(n^2)O(n^{-4}) = O(n^{-2})
\end{aligned}
\tag{4.39}
$$

and

$$
\begin{aligned}
E(T_{8j}) &= \frac{h^{(8)}(\hat{b}_j)}{8!} E(b_j - \hat{b}_j)^8 \\
&= \frac{105}{8!} O(n^{-4}) [-\sum_{i=1}^{n} \hat{m}_{ij}^{(8)}] \\
&= \frac{105}{8!} O(n^{-4}) \left\{ -\sum_{i=1}^{n} \frac{\partial m_{ij}^{(7)}}{\partial \eta_{ij}} (\hat{b}_j) \right\} \\
&= \frac{105}{8!} O(n^{-4}) \left\{ -\sum_{i=1}^{n} \frac{\partial (\partial m_{ij}^{(6)} / \partial \eta_{ij})}{\partial \eta_{ij}} (\hat{b}_j) \right\} \\
&= \frac{105}{8!} O(n^{-4}) \left\{ -\sum_{i=1}^{n} \frac{\partial (\partial [m_{ij}^{(4)}(1 - 12 w_{ij}) - 12 m_{ij}^{(3)2}] / \partial \eta_{ij})}{\partial \eta_{ij}} (\hat{b}_j) \right\} \\
&= \frac{105}{8!} O(n^{-4}) \left\{ -\sum_{i=1}^{n} \frac{\partial [(1 - 2\mu_{ij}) w_{ij} (1 - 60 w_{ij} + 360 w_{ij}^2)]}{\partial \eta_{ij}} (\hat{b}_j) \right\}. \\
&= \frac{105}{8!} O(n^{-4}) \left\{ -n \sum_{i=1}^{n} \hat{w}_{ij} (1 - 126 \hat{w}_{ij} + 1680 \hat{w}_{ij}^2 - 5040 \hat{w}_{ij}^3) / n \right\} \\
&= \frac{105}{8!} O(n^{-4}) O(n) = O(n^{-3})
\end{aligned}
\tag{4.40}
$$

Therefore, the Laplace6 approximation has an error of order $O(n^{-2})$ since

$$
\begin{aligned}
Laplace8 - Laplace6 &= E(T_8) + \frac{1}{2} E(T_4^2) + E(T_3 T_5) \\
&= O(n^{-3}) + O(n^{-2}) + O(n^{-2}) = O(n^{-2})
\end{aligned}
\tag{4.41}
$$

The error of the G-point adaptive Gauss-Hermite quadrature approximation to the integral (4.35) was proved by Liu and Pierce (1994) to be of order $O(n^{-[G/3+1]})$ where [x] is the largest integer not exceeding x. I used their approach to derive the asymptotic behavior of the adaptive Gaussian method and found a slightly different result. Since the

proof is instructive and slightly different from theirs (and hence shows slightly different results), detailed steps of the proof are given here.

In order to apply Liu and Pierce's (1994) approach, $l(b)$ in (4.36) must first be shown to be a unimodal function. To show this, since it is constant with respect to b, let's assume $x_{ij}'\beta = 0$ for simplicity. Then,

$$\begin{aligned} l(b_j) &= \frac{1}{n}\left[\sum_i y_{ij}\eta_{ij} + \sum_i \log(1-\mu_{ij}) - \frac{b_j^2}{2D}\right] \\ &= \frac{\sum_i y_{ij} b_j}{n} - \frac{n\log(1+\exp[b_j])}{n} - \frac{b_j^2}{2nD} \\ &= \bar{y}_j b_j - \log(1+\exp[b_j]) - \frac{b_j^2}{2nD} \end{aligned} \tag{4.42}$$

whence

$$l'(b_j) = \bar{y}_j - \frac{b_j}{nD} - \frac{\exp[b_j]}{1+\exp[b_j]} = 0. \tag{4.43}$$

Since the two terms involving b_j are strictly monotonic in the same direction, this can only have one solution (root). Hence, $l(b_j)$ is unimodal.

For adaptive Gauss-Hermite, (4.35) can be written as

$$\int_{-\infty}^{\infty} f(b)\phi(b;\hat{\mu},\hat{\sigma})db \tag{4.44}$$

where $\phi(.;\mu,\sigma)$ is the normal density with mean μ and standard deviation σ, and

$$\hat{\mu} = \hat{b} \quad (\textit{conditional mode}), \quad \hat{\sigma} = \sqrt{\frac{1}{-h''(\hat{b})}} = \sqrt{\frac{1}{-n\, l''(\hat{b})}} \; . \tag{4.45}$$

Then,

$$f(b) = \frac{g(b)}{\phi(b;\hat{\mu},\hat{\sigma})} = \frac{\exp[nl(b)]}{\frac{1}{\sqrt{2\pi}\hat{\sigma}} \exp\left[-\frac{1}{2}\left(\frac{b-\hat{\mu}}{\hat{\sigma}}\right)^2\right]} = \sqrt{\frac{2\pi}{-nl''(\hat{b})}} \exp[nl(b) - \frac{nl''(\hat{b})}{2}(b-\hat{b})^2] \; . \tag{4.46}$$

Thus,

$$f(\hat{\mu}) = f(\hat{b}) = \sqrt{2\pi}\ \hat{\sigma}\ \exp[nl(\hat{b})] = \sqrt{\frac{2\pi}{-nl''(\hat{b})}} \exp[nl(\hat{b})] \; . \tag{4.47}$$

A Taylor-series expansion of $f(b)$ gives

$$f(b) = f(\hat{b})\pi(b) = f(\hat{b})\ [1 + \sum_{k=1}^{\infty} c_k (b-\hat{b})^k] \tag{4.48}$$

whence

$$\pi(b) = \exp[nl(b) - nl(\hat{b}) - \frac{nl''(\hat{b})}{2}(b-\hat{b})^2]. \tag{4.49}$$

Thus,

$$\begin{aligned}
\int_{-\infty}^{\infty} \exp[nl(b)]db &= \int_{-\infty}^{\infty} f(b)\phi(b;\hat{b},\hat{\sigma})db \\
&= f(\hat{b})\int_{-\infty}^{\infty}[1+\sum_{k=1}^{\infty} c_k(b-\hat{b})^k]\phi(b;\hat{b},\hat{\sigma})db \\
&= f(\hat{b})[1+\sum_{k=1}^{\infty} c_k \int_{-\infty}^{\infty} t^k \phi(t;0,\hat{\sigma})dt] \\
&= f(\hat{b})[1+\sum_{k=1}^{2G-1} c_k \int_{-\infty}^{\infty} t^k \phi(t;0,\hat{\sigma})dt + \sum_{k=2G}^{\infty} c_k \int_{-\infty}^{\infty} t^k \phi(t;0,\hat{\sigma})dt] \ .
\end{aligned} \tag{4.50}$$

The first *2G-1* terms in the expansion (4.49) for $f(b)$ are picked up exactly by the *G*-order (adaptive) Gauss-Hermite quadrature (see Chapter 3) and so the error is of the same order as the integral of the term involving c_{2G}. Now, (by Taylor's theorem)

$$c_{2G} = \frac{1}{(2G)!}\left\{\frac{d^{2G}}{d\,b^{2G}}\pi(b)\Big|_{b=\hat{b}}\right\}. \tag{4.51}$$

Liu (1993) showed that

$$\frac{d^{2G}}{d\,b^{2G}}\pi(b)\Big|_{b=\hat{b}} = O(n^{[2G/3]}) \tag{4.52}$$

where [x] is the largest integer not exceeding x. So, we have

$$\begin{aligned}
c_{2G}\int_{-\infty}^{\infty} t^{2G}\phi(t;0,\hat{\sigma})dt &= c_{2G}\hat{\sigma}^{2G}\int_{-\infty}^{\infty} x^{2G}\phi(x;0,1)dx \\
&= c_{2G}[-n\ l''(\hat{b})]^{-G}\frac{(2G)!}{2^G G!}
\end{aligned} \tag{4.53}$$

the latter fraction being the standard $2G$-th normal moment (see Evans et al., 1993). Thus, (4.54) becomes

$$\frac{1}{(2G)!}O(n^{[2G/3]})[-n\ l''(\hat{b})]^{-G}\frac{(2G)!}{2^G G!} = \frac{[-l''(\hat{b})]^{-G}}{2^G G!}O(n^{[2G/3]-G}) = O(n^{[-G/3]}) \quad \textbf{(4.54)}$$

i.e., the error of the adaptive Gauss-Hermite quadrature is of order $O(n^{[-G/3]})$. Note that for G=1, which makes the adaptive Gauss Hermite the same as standard Laplace (PQL), the error is $O(n^{-1})$ as shown by Tierney and Kadane (1986). Also, for the adaptive Gauss to have an error of the same order as Laplace6 as obtained in (4.42), the smallest G has to be is 4. This seems to corroborate the corresponding errors obtained in chapter 3 for Laplace6 and adaptive Gauss with 4 quadrature points.

Chapter 5

EVALUATION OF METHODS USING SIMULATED DATA

Introduction

This chapter compares four methods: PQL (Raudenbush, 1993), 6th order Laplace (Laplace6) (Yang, 1998), Gauss-Hermite Quadrature (using MIXOR, Hedeker and Gibbons, 1994), Adaptive Gauss-Hermite (using SAS PROC NLMIXED, 1999) using groups of datasets simulated under the same models but varying parameter values and cluster sizes. The simulations are based on two models, both simple hierarchical logistic models, one with a univariate random effect and the other with bivariate ones. The methods are compared in terms of 1) bias of their estimates; 2) mean squared error of the estimates; 3) standard deviation of estimates across replications (datasets); 4) average of standard errors of estimates from method; and 5) computational efficiency (speed).

Univariate Random Effect Model

Simulation Design

Eight groups of datasets were simulated using the following univariate random effect model and specifications. The structure of the datasets follows Rodriguez and Goldman's (1995) rectangular structure. The level-1 and level-2 models are given by

Level-1:

$$\eta_{ij} = \log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_{0j} + \beta_{1j} * (child\ cov)_{ij} \qquad \textbf{(5.1)}$$

Level-2:

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01} * (school\ cov)_j + u_{0j}, \quad u_{0j} \sim N(0,\tau) \\ \beta_{1j} &= \gamma_{10}\end{aligned} \tag{5.2}$$

resulting in the combined model

$$\eta_{ij} = \gamma_{00} + \gamma_{01} * (school\ cov)_j + \gamma_{10} * (child\ cov)_{ij} + u_{0j}, \quad u_{0j} \sim N(0,\tau) \tag{5.3}$$

The model can be thought of as a nested model where the dichotomous outcome y_{ij} is predicted, via the logit link, by a child-level covariate ("child cov") and a school-level covariate ("school cov"), where intercepts but not slopes vary across schools. The level-1 predictor, child cov, was sampled from a normal distribution with mean .0955621 and variance .0676, while school cov, the level-2 covariate, was sampled from a normal distribution with mean -.6857591 and variance .2304. The values of the two fixed effect parameters γ_{01} and γ_{10} are preset to 1. The parameter values set for γ_{00} correspond to small and large values for the conditional expectation

$$\mu_{ij}^{(0)} = E(y_{ij} | u_j = 0) = \frac{1}{1 + \exp[-\eta_{ij}^{(0)}]} \quad \text{where}$$

$$\eta_{ij}^{(0)} = \gamma_{00} + 1 * (-.6857591) + 1 * (.0955621). \tag{5.4}$$

To be exact, the two values used in the simulation for γ_{00}, -1.62 and 0.6653, correspond to conditional expectation values of $\mu_{ij}^{(0)} = 0.0988$ and $\mu_{ij}^{(0)} = 0.5188$, respectively.

That is, the expectation that child i in school j obtains a 1 for y_{ij} is either .1 or .5 (10 or 50%). For the cluster variance parameter τ, two values, 0.25 representing a small cluster variance (i.e., variation among schools) and 1.0 representing a large one, were used in the simulations.

For each of the 4 ($\gamma_{00} \times \tau$) parameter combinations, two groups of 100 datasets were simulated. This results in 8 sets of 100 simulated datasets. In the first group, the datasets consist of 200 clusters (i.e., schools) with a cluster size of 20 children nested in each school and in the second there are 2 children nested within each of 200 schools. The following table summarizes the specifications used in the simulation.

Table 5.1 - Parameter Specifications

τ	μ	γ_{00}	Cluster size	
1.00	0.52	0.6653	20	2
	0.1	-1.62	20	2
0.25	0.52	0.6653	20	2
	0.1	-1.62	20	2

Choice of Number of Replications

The number of replications *N* of simulated datasets was chosen in such a way that a medium effect size (d=0.5) representing the bias of an estimate for a single parameter would be detected with at least 90% power and at 0.05 level of significance α. Table 2.3.5

(p.37) of Statistical Power Analysis for the Behavioral Sciences (Cohen, 1988) gives *sample size*=100 associated with 94% power for a two sample *t* test. So I chose the number of replications *N* of simulated datasets to be 100.

For estimate $\hat{\theta}$ of parameter θ, consider the hypotheses

$$H_0: bias(\hat{\theta})=0 \quad vs \quad H_A: bias(\hat{\theta})\neq 0. \tag{5.5}$$

Since effect size $d = bias(\hat{\theta})/\sigma = (E(\hat{\theta}) - \theta)/\sigma$, this translates to tolerating bias of up to 0.5σ. Note that to compute the *sample* bias, I took the mean of $\hat{\theta}_k - \theta$, where the true value of the parameter θ is known by (simulation) design. Thus, I can compute the estimate of σ as the standard deviation of $\hat{\theta}_k - \theta$. Since PQL underestimates τ for large τ, I am most interested in the bias of the estimation of $\theta = \tau$ for large τ. I already had 100 datasets generated using a bivariate random effects model with (intercept) random effects variance τ=1.625 (and $\gamma_{00} = -1.2$) and I have already found the Laplace6 (L6) and Adaptive Gaussian quadrature (AGQ) estimates of the parameters for these datasets. I also computed the sample standard deviations of the error of estimation. As the standard deviations from the two methods are close to each other, I chose the L6 standard deviations (which are consistently smaller) to compute how large a bias I am willing to tolerate. The results are summarized in the following table. Once two biases (i.e., biases of two methods) are found to be significant in the same direction (i.e., both

Table 5.2 - Percent Tolerated Bias of Estimates for Parameters

Parameter	$\hat{\sigma}$	tolerated bias=1/2($\hat{\sigma}$)	% tolerated bias
$\gamma_{00} = -1.2$	0.111	0.056	4.67%
$\gamma_{01} = 1$	0.104	0.052	5.2%
$\gamma_{10} = 1$	0.068	0.034	3.4%
$\tau = 1.625$	0.291	0.146	8.98%

positive or negative), then they can be compared to see whether one is significantly more biased than the other.

Using the mean square error (MSE), we have the hypotheses

$$H_0: E(\hat{\theta}_1-\theta)^2=E(\hat{\theta}_2-\theta)^2 \quad vs \quad H_A: E(\hat{\theta}_1-\theta)^2\neq E(\hat{\theta}_2-\theta)^2 \tag{5.6}$$

where $\hat{\theta_1}$ refers to L6 estimator and $\hat{\theta_2}$ refers to AGQ estimator. To test this, a (matched) paired t test is used. Define

$$X_k=(\hat{\theta}_{1k}-\theta)^2, \quad Y_k=(\hat{\theta}_{2k}-\theta)^2, \quad d_k=X_k-Y_k \ , \ k=1,...,N. \tag{5.7}$$

By the central limit theorem (CLT), asymptotically (i.e., as N goes to infinity), it is assumed that

$$\bar{d}\sim N(\delta,\sigma^2/N) \ where \ \delta=E(\hat{\theta}_1-\theta)^2-E(\hat{\theta}_2-\theta)^2. \tag{5.8}$$

Alternatively, we can take $d_k = \log(\hat{\theta}_{1k} - \theta)^2 - \log(\hat{\theta}_{2k} - \theta)^2$. (The bar chart of this d_k, for $\theta = \tau$, looks more bell-shaped than that of the original d_k which is clearly skewed to the left.) To test the above hypothesis, $t = \bar{d} / s_{\bar{d}} = \sqrt{N}\,\bar{d} / s$ is used.

A medium effect size (d=0.5) using this paired t test leads to

$$d=\frac{|\mu_1-\mu_2|}{\sigma}=\frac{|MSE_1-MSE_2|}{\sigma}=0.5 \Rightarrow |MSE_1-MSE_2|=0.5\sigma=0.5(.0215)=.0108 \qquad \textbf{(5.9)}$$

where $\hat{\sigma} = .0215$ is the (sample) standard deviation of $X_k - Y_k$ obtained from the (bivariate random effects model) simulation study mentioned above. Taking the difference of the logs and using the delta method with $X_m = (\hat{\theta}_m - \theta)^2$, we have

$$\begin{aligned} E(d_k)&=E[\log(X_{1k})-\log(X_{2k})]=E[\log(\hat{\theta}_{1k}-\theta)^2-\log(\hat{\theta}_{2k}-\theta)^2] \\ &\approx\log[E(\hat{\theta}_{1k}-\theta)^2]-\log[E(\hat{\theta}_{2k}-\theta)^2]=\log\left(\frac{MSE_1}{MSE_2}\right). \end{aligned} \qquad \textbf{(5.10)}$$

Thus

$$d=\frac{|\mu_1-\mu_2|}{\sigma}=\frac{|\log\ MSE_1-\log\ MSE_2|}{\sigma}=0.5 \Rightarrow |\log(\frac{MSE_1}{MSE_2})|=0.5\sigma=0.5(.3866)=.1933 \qquad \textbf{(5.11)}$$

where $\hat{\sigma} = .3866$ is the (sample) standard deviation of $\log X_{1k} - \log X_{2k}$ computed from the simulation study mentioned above. This implies *relative efficiency of method* 2

to method 1 = $MSE_1 / MSE_2 = e^{.1933} = 1.213$. The relative efficiencies for the MSE's of fixed effects estimates are similarly computed. Table 5.3 summarizes the results where $\hat{\sigma}$ for each parameter is the standard deviation of $\log X_{1k} - \log X_{2k}$ computed for each parameter from the simulation study mentioned above. I will take these relative efficiencies as the largest ones I am willing to tolerate before I must declare one MSE is larger than another with adequate power.

Table 5.3 - Relative Efficiencies for MSE's of Estimates

Parameter	$\hat{\sigma}$	$rel.eff. = \exp(\hat{\sigma} / 2)$
γ_{00}	0.495	1.281
γ_{01}	1.308	1.923
γ_{10}	0.857	1.535
τ	0.387	1.213

Running the Algorithms

PQL (which is similar to the standard Laplace) and Laplace6 were implemented in Raudenbush, Bryk and Congdon's HLM program. So the HLM program Version 5.20 was used to run these two methods with the specification that Laplace6 is going to be run following PQL. The non-adaptive Gauss-Hermite method was run using Hedeker and Gibbons' Mixed Ordinal Regression Model (MIXOR) Version 2.0 program. For the numerical integration that is required in this method, the default number of quadrature points set by MIXOR for a single random effect, namely 10, was chosen. The adaptive

Gauss-Hermite method was run on SAS Version 7 using PROC NLMIXED by Wolfinger. This procedure (algorithm) selects the number of quadrature points adaptively. The number selected is the one that gives a likelihood value of a negligible difference if the next higher number of quadrature points was used. In this simulation study, it turns out that the number of quadrature points selected by the procedure for all the runs was either 3 or 5. Also, no initial values were specified for the parameters so that the procedure itself assigns the default initial value of 1 to each parameter. All programs were run on a 450 MHz PC with Pentium II processor.

In some of the runs, especially when the cluster size is small (cluster size=2), MIXOR had problems giving results. These problems include terminating but not giving estimates at all, estimating only fixed effects parameters (i.e., estimating a different model), not converging in 10000 iterations (after which I manually stopped it using CTRL-C) and giving unreasonable estimate values (especially for τ). All these cases, where MIXOR didn't run properly, were not considered and comparison with the other methods should be based on those cases (datasets) where MIXOR ran properly.

Results

Bias: large cluster size. A cursory look at Table 4 (% Bias of Estimates) reveals that when the cluster size is large (20 in this case), the biases of all the estimates from the three methods except PQL are within reasonable limits. The only exception is for γ_{10} when both the conditional expectation (corresponding to $\gamma_{00} = -1.62$) and τ ($\tau = 0.25$) are small, in which case none of the biases of the estimates are within reasonable limits. For PQL, only the bias of small τ ($\tau = 0.25$) when the conditional

Table 5.4 - Percent Bias of Estimates

Cluster size	Parameter	PQL	Laplace6	Gauss	AGQ
20	γ_{00} = .6653	-3.13%	2.49%	2.06%	2.74%
	γ_{01} = 1	-3.91%	1.48%	1.01%	1.72%
	γ_{10} = 1	-2.73%	1.57%	1.66%	1.65%
	τ = 1.0	-12.70%	-1.13%	-1.10%	-0.81%
	γ_{00} = .6653	-2.03%	0.95%	0.95%	0.92%
	γ_{01} = 1	-2.44%	0.49%	0.50%	0.47%
	γ_{10} = 1	-1.25%	1.52%	1.53%	1.52%
	τ = 0.25	-7.44%	0.96%	1.00%	0.72%
	γ_{00} = -1.62	-5.80%	0.27%	0.28%	0.23%
	γ_{01} = 1	-10.06%	-2.38%	-2.08%	-2.05%
	γ_{10} = 1	-6.01%	-2.01%	-2.03%	-2.06%
	τ = 1.0	-24.60%	-6.65%	-6.63%	-6.82%
	γ_{00} = -1.62	-4.57%	-1.72%	-1.73%	-1.73%
	γ_{01} = 1	-0.21%	2.79%	2.80%	2.80%
	γ_{10} = 1	-6.29%	-4.50%	-4.51%	-4.51%
	τ = 0.25	-16.04%	-4.60%	-4.88%	-4.88%
2	γ_{00} = .6653	-9.79% -10.09% (N=98)	8.37% 8.45% (N=98)	6.82% (N=98)	6.55% 6.57% (N=98)
	γ_{01} = 1	-15.31% -15.22% (N=98)	1.57% 2.00% (N=98)	0.61% (N=98)	-0.04% 0.36% (N=98)
	γ_{10} = 1	-18.49% -17.71% (N=98)	-2.32% -1.21% (N=98)	-2.39% (N=98)	-3.71% -2.62% (N=98)
	τ = 1.0	-62.66% -61.91% (N=98)	19.23% 21.66% (N=98)	3.54% (N=98)	-0.22% 1.82% (N=98)
	γ_{00} = .6653	-8.40% -8.09% (N=79)	-2.27% -0.30% (N=79)	-0.56% (N=79)	-2.74% -0.60% (N=79)
	γ_{01} = 1	-5.68% -5.47% (N=79)	0.52% 2.38% (N=79)	2.18% (N=79)	0.71% 2.14% (N=79)
	γ_{10} = 1	1.95% 2.48% (N=79)	8.05% 10.21% (N=79)	10.21% (N=79)	8.87% 10.19% (N=79)
	τ = 0.25	-47.84% -34.36% (N=79)	33.52% 68.60% (N=79)	59.04% (N=79)	25.24% 58.52% (N=79)
	γ_{00} = -1.62	-13.93% -13.87% (N=88)	1.33% 3.27% (N=88)	3.96% (N=88)	-1.96% -0.47% (N=88)
	γ_{01} = 1	-4.65% -5.09% (N=88)	7.60% 8.64% (N=88)	9.60% (N=88)	5.66% 6.48% (N=88)
	γ_{10} = 1	-11.14% -12.09% (N=88)	0.94% 1.47% (N=88)	0.78% (N=88)	-1.46% -1.83% (N=88)
	τ = 1.0	-60.85% -56.05% (N=88)	4.99% 17.99% (N=88)	25.76% (N=88)	-17.91% -7.74% (N=88)
	γ_{00} = -1.62	-2.85% -2.24% (N=49)	4.07% 11.42% (N=49)	11.92% (N=49)	2.78% 8.55% (N=49)
	γ_{01} = 1	-3.46% -3.31% (N=49)	0.47% 4.36% (N=49)	6.44% (N=49)	-0.40% 4.35% (N=49)
	γ_{10} = 1	-9.41% -3.61% (N=49)	-4.82% 5.53% (N=49)	5.17% (N=49)	-3.72% 3.44% (N=49)
	τ = 0.25	-36.16% 24.24% (N=49)	61.92% 216.76% (N=49)	242.24% (N=49)	29.04% 154.72% (N=49)

expectation is large (corresponding to $\gamma_{00} = .6653$) is within limits. All other PQL estimates of τ are significantly underestimated (negatively biased). The negative bias of PQL estimate for a large τ, along with a small conditional expectation, resulted in the negative bias (underestimation) of all the other parameters (i.e., the fixed effects). For a small conditional expectation, γ_{10} was still underestimated by PQL even for a small τ.

Bias: small cluster size. When the cluster size is small (2 in this case), GQ (using MIXOR) had difficulty estimating the parameters from some of the data, especially when the dataset was generated with small τ ($\tau = 0.25$). When both τ and conditional expectation are large, GQ worked (MIXOR ran) well on all but two of the datasets; in one case, it terminated but didn't give estimates, in the other, it fitted a model without the random effect. In this case (of large τ and conditional expectation), the methods based on Gaussian quadratures worked the best, only the bias of the estimate of γ_{00} being significant, and PQL did the worst, the estimates of all parameters being significantly negatively biased; for Laplace6, only the biases of γ_{00} and τ are significant, both being positively biased.

When both τ and conditional expectation are small, GQ (MIXOR) performed the worst; first of all, giving reasonable results in only 49 of the 100 cases (datasets), and secondly, even in those cases giving a significant positive bias for all estimates (especially for τ, where the per cent bias is 242%). In this case, PQL gave the least bias for τ with 24% (though it still was significantly biased). The fixed effects were all well estimated, with only γ_{10} being estimated with a larger than tolerable bias. For AGQ and

Laplace6, the comparable estimates of γ_{00} were biased, as well (i.e., besides γ_{10}). It should be pointed out that when all 100 replications (datasets) were considered (instead of just the 49 for which MIXOR ran properly), the per cent biases of both Laplace6 and AGQ for γ_{00} were no longer larger than the percent tolerated bias. This is important because the 'bad' cases for MIXOR are actually better for others.

When τ is small but conditional expectation is large, PQL's estimate has tolerable bias only for γ_{10}, while the biases from all the other methods (for the 79 cases for which MIXOR gave reasonable estimates) were significant for γ_{10} and τ. When τ is large but conditional expectation is small, PQL's estimate has tolerable bias only for γ_{01}, while the biases from all the other methods were significant for γ_{01} and τ. (For the 88 cases where MIXOR gave reasonable results, the percent bias of the AGQ estimate for τ was less than the tolerated.)

Mean Squared Error: large cluster size. From the Mean Squared Errors table (Table 5.5), we notice that when the cluster size is large, the MSE of the PQL estimate for large τ is larger than the other MSEs, while for small τ, only the MSE of the PQL estimate of γ_{00} corresponding to a small conditional expectation is larger than the others. MSEs for GQ, AGQ and Laplace6 are remarkably similar.

Mean Squared Error: small cluster size. When the cluster size is small, the MSEs of the PQL estimates of small τ are much smaller than the corresponding MSEs of all the other estimates while the MSEs of the AGQ estimates of a large τ are the smallest in their

Table 5.5 - Mean Squared Errors

Cluster size	Parameter	PQL	Laplace6	Gauss	AGQ
20	$\gamma_{00} = .6653$	0.0191	0.0212	0.0214	0.0214
	$\gamma_{01} = 1$	0.0337	0.0363	0.0381	0.0367
	$\gamma_{10} = 1$	0.0209	0.0223	0.0223	0.0223
	$\tau = 1.0$	0.0296	0.0187	0.0185	0.0190
	$\gamma_{00} = .6653$	0.0075	0.0078	0.0078	0.0078
	$\gamma_{01} = 1$	0.0134	0.0137	0.0137	0.0137
	$\gamma_{10} = 1$	0.0127	0.0134	0.0134	0.0134
	$\tau = 0.25$	0.0017	0.0017	0.0017	0.0016
	$\gamma_{00} = -1.62$	0.0301	0.0252	0.0256	0.0252
	$\gamma_{01} = 1$	0.0420	0.0376	0.0375	0.0376
	$\gamma_{10} = 1$	0.0383	0.0382	0.0382	0.0382
	$\tau = 1.0$	0.0773	0.0341	0.0349	0.0348
	$\gamma_{00} = -1.62$	0.0141	0.0101	0.0101	0.0101
	$\gamma_{01} = 1$	0.0173	0.0195	0.0195	0.0195
	$\gamma_{10} = 1$	0.0439	0.0435	0.0435	0.0436
	$\tau = 0.25$	0.0058	0.0061	0.0060	0.0060
2	$\gamma_{00} = .6653$	0.0428 0.0434 (N=98)	0.0608 0.0618 (N=98)	0.0582 (N=98)	0.0573 0.0582 (N=98)
	$\gamma_{01} = 1$	0.0746 0.0753 (N=98)	0.0771 0.0778 (N=98)	0.0747 (N=98)	0.0740 0.0747 (N=98)
	$\gamma_{10} = 1$	0.2158 0.2130 (N=98)	0.2861 0.2848 (N=98)	0.2718 (N=98)	0.2725 0.2707 (N=98)
	$\tau = 1.0$	0.4147 0.4029 (N=98)	0.7228 0.7173 (N=98)	0.2202 (N=98)	0.2175 0.2016 (N=98)
	$\gamma_{00} = .6653$	0.0447 0.0474 (N=79)	0.0522 0.0569 (N=79)	0.0556 (N=79)	0.0522 0.0555 (N=79)
	$\gamma_{01} = 1$	0.0596 0.0617 (N=79)	0.0721 0.0776 (N=79)	0.0752 (N=79)	0.0695 0.0750 (N=79)
	$\gamma_{10} = 1$	0.1670 0.1822 (N=79)	0.1984 0.2218 (N=79)	0.2218 (N=79)	0.1983 0.2215 (N=79)
	$\tau = 0.25$	0.0307 0.0228 (N=79)	0.2152 0.2563 (N=79)	0.1455 (N=79)	0.1255 0.1422 (N=79)
	$\gamma_{00} = -1.62$	0.1200 0.1269 (N=88)	0.1312 0.1404 (N=88)	0.1501 (N=88)	0.1100 0.1164 (N=88)
	$\gamma_{01} = 1$	0.1134 0.1252 (N=88)	0.1504 0.1666 (N=88)	0.1679 (N=88)	0.1388 0.1538 (N=88)
	$\gamma_{10} = 1$	0.4074 0.4475 (N=88)	0.5641 0.6248 (N=88)	0.6160 (N=88)	0.5142 0.5663 (N=88)
	$\tau = 1.0$	0.4362 0.3689 (N=88)	0.5060 0.4596 (N=88)	0.7498 (N=88)	0.3543 0.2831 (N=88)
	$\gamma_{00} = -1.62$	0.0684 0.0557 (N=49)	0.1030 0.1273 (N=49)	0.1436 (N=49)	0.0914 0.1028 (N=49)
	$\gamma_{01} = 1$	0.1242 0.1171 (N=49)	0.1350 0.1370 (N=49)	0.1480 (N=49)	0.1353 0.1386 (N=49)
	$\gamma_{10} = 1$	0.3351 0.3299 (N=49)	0.3718 0.4056 (N=49)	0.4131 (N=49)	0.3757 0.3921 (N=49)
	$\tau = 0.25$	0.0537 0.0511 (N=49)	0.3057 0.5679 (N=49)	0.7948 (N=49)	0.1904 0.3285 (N=49)

groups. As far as the estimation of τ is concerned, the PQL and AGQ estimates seem to have the lowest MSEs of the four methods.

For small τ and small conditional expectation, the order of the MSEs for τ estimates from smallest to largest was PQL, AGQ, Laplace6, GQ, each being significantly larger than its predecessor in terms of relative efficiency (according to my criteria in Table 5.3). For this case, the MSE of the PQL estimate of γ_{00} was the smallest followed by that of AGQ which is significantly smaller than that of the largest, GQ. The Laplace6 estimate was not significantly different in terms of relative efficiency from either of the Gauss based estimates.

For large τ and large conditional expectation, the MSE of the Laplace6 estimate of τ was by far the largest followed by that of PQL which was still significantly larger than the other two MSE's that are based on Gaussian quadratures. For this case, the MSE of the PQL estimate of γ_{00} was significantly smaller than the other MSEs.

When τ is large but the conditional expectation is small, the MSE of the GQ estimate was by far the largest while that of the AGQ estimate was by far the smallest; there was no significant difference in terms of relative efficiency between the MSEs of PQL and Laplace6.

For small τ and large conditional expectation, the MSE of the PQL estimate of τ was by far the smallest, followed by those of the AGQ and GQ estimates in that order, both of which were significantly smaller than the MSE of Laplace6 but were not significantly different from each other. The MSEs of the fixed effects estimates from the four methods were not different from each other.

Accuracy of Standard Error Estimates. For each method, the averages over all replications of the printed standard errors of the estimates were computed and were evaluated as estimators of the true standard deviations of the estimates. The latter were estimated by the standard deviations across replications of the estimates. (Note that the squares of these are the unbiased estimates of the variances.) These two standard error

Table 5.6 - Standard Error Estimates for PQL

Cluster size	Parameter	Avg SE of Estimates	SD of Estimates	Root MSE
20	γ_{00} = .6653	0.1330	0.1371	0.1382
	γ_{01} = 1	0.1588	0.1802	0.1836
	γ_{10} = 1	0.1385	0.1426	0.1446
	τ = 1.0	0.1124	0.1166	0.1720
	γ_{00} = .6653	0.0857	0.0858	0.0866
	γ_{01} = 1	0.1023	0.1138	0.1174
	γ_{10} = 1	0.1309	0.1124	0.1127
	τ = 0.25	0.0451	0.0367	0.0412
	γ_{00} = -1.62	0.1375	0.1466	0.1735
	γ_{01} = 1	0.1733	0.1794	0.2049
	γ_{10} = 1	0.1894	0.1873	0.1957
	τ = 1.0	0.1267	0.1303	0.2780
	γ_{00} = -1.62	0.0996	0.0935	0.1187
	γ_{01} = 1	0.1308	0.1322	0.1315
	γ_{10} = 1	0.1953	0.2010	0.2095
	τ = 0.25	0.0707	0.0648	0.0762
2	γ_{00} = .6653	0.2056	0.1972	0.2069
	γ_{01} = 1	0.2429	0.2273	0.2731
	γ_{10} = 1	0.4213	0.4283	0.4645
	τ = 1.0	0.2519	0.1491	0.6440
	γ_{00} = .6653	0.1969	0.2049	0.2114
	γ_{01} = 1	0.2348	0.2386	0.2441
	γ_{10} = 1	0.4138	0.4103	0.4087
	τ = 0.25	0.2276	0.1287	0.1752
	γ_{00} = -1.62	0.2543	0.2642	0.3464
	γ_{01} = 1	0.3372	0.3352	0.3367
	γ_{10} = 1	0.5943	0.6317	0.6383
	τ = 1.0	0.4626	0.2580	0.6605
	γ_{00} = -1.62	0.2644	0.2587	0.2615
	γ_{01} = 1	0.3543	0.3525	0.3524
	γ_{10} = 1	0.6290	0.5740	0.5789
	τ = 0.25	0.4956	0.2144	0.2317

estimates were computed and tabulated along with the root MSE and discussed for each method.

For PQL (Table 5.6), the averages of standard errors of estimates appear to be quite good estimates of the corresponding standard deviations of estimates when the cluster size is large. But when cluster size is small, the averages of the standard errors of the τ estimates consistently overestimate (are larger than) the standard deviations of the estimates. For all these cases, except for small τ along with small conditional expectation, there were large discrepancies between the standard deviations of estimates and root MSEs, indicating significant biases which were confirmed by the Bias table (Table 5.4). The discrepancies (and the biases) were also large for large τ, when cluster size was large. As well, there appeared to be discrepancies between the standard deviations and root MSEs for the intercept for small conditional expectation.

The averages of standard errors of estimates are also quite close for Laplace6 (Table 5.7) to their corresponding standard deviations of estimates, when the cluster size is large. When cluster size is small, the average standard errors of estimates of τ appear to underestimate the standard deviations when the conditional expectation is large and overestimate the standard deviations when the conditional expectation is small. For small conditional expectation, the average standard error of the child slope γ_{10} and its corresponding standard deviation of estimates appear to be somewhat discrepant. In general, there doesn't appear to be that much of a discrepancy between the standard deviations of estimates and the root MSEs for Laplace6.

Table 5.7 - Standard Error Estimates for Laplace6

Cluster size	Parameter	Avg SE of Estimates	SD of Estimates	Root MSE
20	γ_{00} = .6653	0.1414	0.1453	0.1456
	γ_{01} = 1	0.1696	0.1910	0.1905
	γ_{10} = 1	0.1430	0.1491	0.1493
	τ = 1.0	0.1405	0.1369	0.1367
	γ_{00} = .6653	0.0896	0.0886	0.0883
	γ_{01} = 1	0.1086	0.1173	0.1170
	γ_{10} = 1	0.1355	0.1155	0.1158
	τ = 0.25	0.0520	0.0409	0.0412
	γ_{00} = -1.62	0.1535	0.1594	0.1587
	γ_{01} = 1	0.1930	0.1934	0.1939
	γ_{10} = 1	0.1984	0.1954	0.1954
	τ = 1.0	0.1729	0.1731	0.1847
	γ_{00} = -1.62	0.1079	0.0971	0.1005
	γ_{01} = 1	0.1399	0.1375	0.1396
	γ_{10} = 1	0.2019	0.2048	0.2086
	τ = 0.25	0.0823	0.0774	0.0781
2	γ_{00} = .6653	0.2597	0.2414	0.2466
	γ_{01} = 1	0.3107	0.2785	0.2777
	γ_{10} = 1	0.4967	0.5371	0.5349
	τ = 1.0	0.6097	0.8323	0.8502
	γ_{00} = .6653	0.2175	0.2291	0.2285
	γ_{01} = 1	0.2630	0.2698	0.2685
	γ_{10} = 1	0.4426	0.4403	0.4454
	τ = 0.25	0.3751	0.4586	0.4639
	γ_{00} = -1.62	0.3624	0.3633	0.3622
	γ_{01} = 1	0.4178	0.3823	0.3878
	γ_{10} = 1	0.6791	0.7548	0.7511
	τ = 1.0	0.9309	0.7132	0.7113
	γ_{00} = -1.62	0.3299	0.3157	0.3209
	γ_{01} = 1	0.3907	0.3692	0.3674
	γ_{10} = 1	0.6699	0.6109	0.6098
	τ = 0.25	0.7158	0.5334	0.5529

For Gaussian quadratures (Table 5.8), only the average of standard errors of τ estimates for small τ and small conditional expectation appear to differ much from (underestimate) the standard deviation of estimates, when the cluster size is large. When the cluster size is small, the average standard errors of τ estimates were always different

Table 5.8 - Standard Error Estimates for Gauss

Cluster size	Parameter	Avg SE of Estimates	SD of Estimates	Root MSE
20	γ_{00} = .6653	0.1387	0.1463	0.1463
	γ_{01} = 1	0.1664	0.1958	0.1952
	γ_{10} = 1	0.1431	0.1490	0.1493
	τ = 1.0	0.1372	0.1363	0.1360
	γ_{00} = .6653	0.0897	0.0886	0.0883
	γ_{01} = 1	0.1086	0.1174	0.1170
	γ_{10} = 1	0.1355	0.1155	0.1158
	τ = 0.25	0.0263	0.0409	0.0412
	γ_{00} = -1.62	0.1527	0.1606	0.1600
	γ_{01} = 1	0.1919	0.1935	0.1936
	γ_{10} = 1	0.1984	0.1955	0.1954
	τ = 1.0	0.1660	0.1756	0.1868
	γ_{00} = -1.62	0.1078	0.0971	0.1005
	γ_{01} = 1	0.1398	0.1375	0.1396
	γ_{10} = 1	0.2019	0.2048	0.2086
	τ = 0.25	0.0403	0.0770	0.0775
2	γ_{00} = .6653	0.2535 (N=98)	0.2382 (N=98)	0.2412 (N=98)
	γ_{01} = 1	0.3036 (N=98)	0.2747 (N=98)	0.2733 (N=98)
	γ_{10} = 1	0.4934 (N=98)	0.5235 (N=98)	0.5213 (N=98)
	τ = 1.0	0.6160 (N=98)	0.4703 (N=98)	0.4693 (N=98)
	γ_{00} = .6653	0.2240 (N=79)	0.2374 (N=79)	0.2358 (N=79)
	γ_{01} = 1	0.2721 (N=79)	0.2751 (N=79)	0.2742 (N=79)
	γ_{10} = 1	0.4535 (N=79)	0.4627 (N=79)	0.4710 (N=79)
	τ = 0.25	0.2632 (N=79)	0.3540 (N=79)	0.3814 (N=79)
	γ_{00} = -1.62	0.3790 (N=88)	0.3842 (N=88)	0.3874 (N=88)
	γ_{01} = 1	0.4315 (N=88)	0.4007 (N=88)	0.4098 (N=88)
	γ_{10} = 1	0.6926 (N=88)	0.7893 (N=88)	0.7849 (N=88)
	τ = 1.0	1.1703 (N=88)	0.8315 (N=88)	0.8659 (N=88)
	γ_{00} = -1.62	0.3967 (N=49)	0.3294 (N=49)	0.3789 (N=49)
	γ_{01} = 1	0.4314 (N=49)	0.3832 (N=49)	0.3847 (N=49)
	γ_{10} = 1	0.7101 (N=49)	0.6473 (N=49)	0.6427 (N=49)
	τ = 0.25	0.8847 (N=49)	0.6610 (N=49)	0.8915 (N=49)

from the standard deviations of estimates (larger except for small τ when the conditional expectation is large). When the conditional expectation was small, the average standard errors of the γ_{10} estimates were somewhat different as well from their standard deviation counterparts. When both τ and the conditional expectation were small, almost all the averages of the standard errors of the estimates (except perhaps the school effect γ_{01})

were different from their corresponding standard deviations. This was the case where Gauss had difficulty giving estimates with only 49 out of the 100 datasets giving sensible results. Only in this case was there a discrepancy between the standard deviation of the (small) τ estimates and their root MSEs. For no other estimate did Gauss appear to have a

Table 5.9 - Standard Error Estimates for AGQ

Cluster size	Parameter	Avg SE of Estimates	SD of Estimates	Root MSE
20	γ_{00} = .6653	0.1410	0.1457	0.1463
	γ_{01} = 1	0.1684	0.1917	0.1916
	γ_{10} = 1	0.1420	0.1493	0.1493
	τ = 1.0	0.1382	0.1382	0.1378
	γ_{00} = .6653	0.0884	0.0886	0.0883
	γ_{01} = 1	0.1056	0.1173	0.1170
	γ_{10} = 1	0.1331	0.1155	0.1158
	τ = 0.25	0.0506	0.0408	0.0400
	γ_{00} = -1.62	0.1514	0.1595	0.1587
	γ_{01} = 1	0.1892	0.1938	0.1939
	γ_{10} = 1	0.1939	0.1953	0.1954
	τ = 1.0	0.1664	0.1745	0.1865
	γ_{00} = -1.62	0.1046	0.0971	0.1005
	γ_{01} = 1	0.1351	0.1375	0.1396
	γ_{10} = 1	0.1974	0.2048	0.2088
	τ = 0.25	0.0786	0.0769	0.0775
2	γ_{00} = .6653	0.2496	0.2365	0.2394
	γ_{01} = 1	0.2977	0.2734	0.2720
	γ_{10} = 1	0.4792	0.5234	0.5220
	τ = 1.0	0.5745 (N=98)	0.4687 0.4509 (N=98)	0.4664 0.4490 (N=98)
	γ_{00} = .6653	0.2028 (N=99)	0.2288	0.2285
	γ_{01} = 1	0.2443 (N=99)	0.2649	0.2636
	γ_{10} = 1	0.4209 (N=99)	0.4386	0.4453
	τ = 0.25	0.4123 (N=79)	0.3503 0.3498 (N=79)	0.3543 0.3771 (N=79)
	γ_{00} = -1.62	0.3229 (N=99)	0.3317	0.3317
	γ_{01} = 1	0.3745	0.3701	0.3726
	γ_{10} = 1	0.6216	0.7205	0.7171
	τ = 1.0	0.7308 (N=91)	0.5705	0.5952
	γ_{00} = -1.62	0.2971 (N=96)	0.3004	0.3023
	γ_{01} = 1	0.3450 (N=95)	0.3696	0.3678
	γ_{10} = 1	0.5368 (N=95)	0.6149	0.6129
	τ = 0.25	0.7263 (N=55)	0.4325	0.4363

discrepancy between the standard deviation and root MSE. Incidentally, I used the delta method to compute the standard error of the random effects variance for GQ, since MIXOR gives the standard error of the random effects standard deviation instead of the variance.

For AGQ (Table 5.9), the average standard errors of estimates and the corresponding standard deviations were always quite close when the cluster size was large. When the cluster size was small, the standard errors of the τ estimates tend to overestimate the corresponding standard deviations. For large τ and small conditional expectation, the standard error of the child effect (slope) γ_{10} estimates also seemed to be discrepant from (underestimate) the corresponding standard deviation. There appeared to be no discrepancy between the standard deviations of estimates and the root MSEs for any of the estimates.

The standard deviations of estimates across replications are consistently the lowest for PQL, especially for the estimation of τ when the cluster size is small. This may be attributable to the larger biases that PQL suffers from compared to the other methods. For the latter cases (estimation of τ when the cluster size is small), AGQ has the next lowest standard deviations of estimates. In view of the fact that AGQ suffers the least biases among all the methods, this suggests that AGQ estimates uncertainty more accurately than do the other methods.

Computational Efficiency (speed). Incredibly, for those runs that converge for GQ, GQ was by far the fastest. But this is an unfair comparison because while the other methods (actually programs) eventually converge for virtually all datasets, GQ went into

an infinite loop for some of them that I had to manually stop (after 10000 iterations). Incidentally, GQ always converged when the cluster size was large (cluster size=20). For large cluster size (n=20), AGQ appeared to be by far the slowest. In this case, PQL is the next fastest followed by Laplace6, which is understandable because, in HLM, PQL has to finish before Laplace6 begins.

Table 5.10 - Average Speed in Seconds

Cluster size	Parameter		PQL[1]	Laplace6[2]	Gauss	AGQ
	τ	γ_{00}				
20	1.0	0.6653	8.58	18.61	2.71	39.72
	0.25	0.6653	8.62	18.75	2.39	43.30
	1.0	-1.62	8.73	21.73	3.11	50.93
	0.25	-1.62	9.43	22.08	2.58	60.45
2	1.0	0.6653	9.73 (N=98)	12.59 (N=98)	0.47 (N=98)	5.34 (N=98)
	0.25	0.6653	13.40 (N=79)	14.37 (N=79)	12.06 (N=79)	19.87 (N=79)
	1.0	-1.62	12.81 (N=88)	17.78 (N=88)	0.69 (N=88)	7.33 (N=88)
	0.25	-1.62	37.59 (N=49)	39.94 (N=49)	0.87 (N=49)	14.69 (N=49)

When the cluster size is small, AGQ performed quite well in terms of speed. It was the next fastest (after GQ) on the average in all cases except when τ was small and the conditional expectation was large, where it was the slowest. Even in this case, it was by far faster than the other two when run over all the datasets, including those that didn't produce sensible results under MIXOR (GQ), some runs of which went into infinite loops for GQ.

Summary of Results. In summary, when the cluster size is large, all the methods appear to have performed quite well except that PQL was frequently the most biased and sometimes had the largest MSEs. However, it (PQL) gave the smallest standard

[1] For PQL and Laplace6, time=creation of SSM file + running of actual model.
[2] For Laplace6, PQL has to be run first so as to get initial estimates so its time is always greater than PQL.

deviations of estimates and average standard errors of estimates (perhaps due to its large bias), though not by that much, and had the next fastest time following GQ. In this case, GQ had the best performance, being by far the fastest among all methods and having comparable biases (non-significant), MSE, standard deviations of estimates and average standard errors of estimates with the other two methods (Laplace6 and AGQ). Laplace6 was more than twice as fast on the average as AGQ and comparable to it in other respects.

When cluster size is small, GQ appeared to be the worst offender, not giving results many times, especially when the random effects variance is small (and also the conditional expectation is small), and when it gave results they were sometimes biased, especially when the random effects variance and/or the conditional expectation are small (when both are small, its estimates were always biased). AGQ appears to be the best in this instance being faster (overall) than the remaining methods, and having the least MSEs. Besides, whenever it was biased, Laplace6 was biased, too, with one case (when both variance τ and conditional expectation are large) where Laplace6 was biased and AGQ was not, and its standard deviations of estimates are almost always smaller than Laplace6. Laplace6 is biased in fewer instances than PQL, especially when the random effects variance is large, but sometimes had more MSE than PQL. PQL had the smallest standard deviations of estimates as well as smallest average standard errors of all methods which is likely due to its having the most bias.

Bivariate Random Effects Model

A group of 100 datasets was simulated using a bivariate random effects model which has the same level-1 equation as the univariate case, but differs in level-2 equations which are now replaced by

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01} * (school\ cov)_j + u_{0j}, \\ \beta_{1j} &= \gamma_{10} + u_{1j} \end{aligned} \tag{5.12}$$

so that the combined model becomes

$$\eta_{ij} = \gamma_{00} + \gamma_{01} * (school\ cov)_j + \gamma_{10} * (child\ cov)_{ij} + u_{0j} + u_{1j} * (child\ cov)_{ij}, \quad \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{bmatrix} \right) \tag{5.13}$$

where $\tau_{00} = 1.625$, $\tau_{11} = 0.25$, and $\tau_{01} = 0.1$.

As in the univariate case, this model can be thought of as a nested model where the dichotomous outcome y_{ij} is predicted by a child-level covariate ("child cov") and a school-level covariate ("school cov"). In this case, slopes *as well as* intercepts vary across schools. The level-1 predictor, child cov, was sampled from a normal distribution with mean .0955621 (and variance 1), while school cov, the level-2 covariate, was sampled from a normal distribution with mean -.6857591 (variance 1). The values of the two fixed effect parameters γ_{01} and γ_{10} are preset to 1. The parameter values for γ_{00} was set at -1.2 which corresponds to an conditional expectation $E(y_{ij} \mid \boldsymbol{u}_j = \mathbf{0})$ of 0.14.

Again the rectangular structure of Rodriguez and Goldman (1995) is followed in the datasets with 20 hypothetical children nested within each of 200 hypothetical schools

for a total of 4000 children in each dataset. The number of replications was determined in a similar manner as in the univariate cases. In fact, it was this group of 100 datasets that was previously generated that was used to obtain estimates for the standard deviations in the formula for effect size.

The four methods were run on the 100 datasets generated with the above hierarchical logistic model with bivariate random effects the same way (i.e., with the same program specifications) as they were for the groups of datasets with univariate random effects and on the same machine. The only difference was that I specified an initial value of -1.0 for the intercept parameter γ_{00} for AGQ (i.e., PROC NLMIXED in SAS). I did this after SAS gave a 'No valid parameter points were found' message and terminated without producing results when I ran the model with no initial (starting) values for the parameters. By default, SAS assigns an initial value of 1 to each parameter. Incidentally, SAS adaptively selected 7 to be the number of quadrature points used for each of the runs. Again, 10 quadrature points were specified for MIXOR.

The results are summarized in Tables 5.11 and 5.12. From Table 5.11 we can see that PQL's estimates of all parameters are (negatively) biased while the percent biases from the other estimates are within reasonable limits. Incidentally, the same percent tolerated biases as for the univariate case are used here for the common parameters while the percent tolerated biases for τ_{11} and τ_{01} are computed to be 18.4% and 49% respectively.

Using the same tolerated relative efficiencies of the MSEs as before for the common parameters and taking the computed tolerated relative efficiencies of 1.689 for τ_{01} and 1.650 for τ_{11}, the MSE of the PQL estimates were found to be larger than the others

Table 5.11 - Percent Bias, MSE and Speed for Bivariate Random Effects Model

Title	Parameter	PQL	Laplace6	Gauss	AGQ
%Bias	γ_{00}=-1.2	-9.33% -9.26% (N=97)	-4.33% -0.38% (N=97)	-0.04% 0.03% (N=97)	-0.18% (N=97)
	γ_{01}=1	-11.26% -11.46% (N=97)	-1.34% -1.56% (N=97)	-0.51% -0.77% (N=97)	-0.51% (N=97)
	γ_{10}=1	-9.01% -9.11% (N=97)	-0.71% -0.82% (N=97)	-0.63% -0.74% (N=97)	-0.80% (N=97)
	τ_{00}=1.625	-23.38% -23.80% (N=97)	-2.06% -2.61% (N=97)	-2.17% -2.63% (N=97)	-1.48% (N=97)
	τ_{01}=0.1	-48.20% -49.00% (N=97)	-8.39% -9.70% (N=97)	-8.11% -9.20% (N=97)	-7.85% (N=97)
	τ_{11}=0.25	-41.16% -40.92% (N=97)	-1.60% -1.28% (N=97)	-4.74% -4.48% (N=97)	-5.05% (N=97)
MSE	γ_{00}=-1.2	0.0223 0.0224 (N=97)	0.0122 0.0125 (N=97)	0.0145 0.0149 (N=97)	0.0126 (N=97)
	γ_{01}=1	0.0212 0.0218 (N=97)	0.0108 0.0110 (N=97)	0.0123 0.0125 (N=97)	0.0112 (N=97)
	γ_{10}=1	0.0116 0.0118 (N=97)	0.0047 0.0047 (N=97)	0.0048 0.0049 (N=97)	0.0048 (N=97)
	τ_{00}=1.625	0.1896 0.1945 (N=97)	0.0847 0.0848 (N=97)	0.0970 0.0983 (N=97)	0.0925 (N=97)
	τ_{01}=0.1	0.0074 0.0074 (N=97)	0.0094 0.0094 (N=97)	0.0102 0.0103 (N=97)	0.0103 (N=97)
	τ_{11}=0.25	0.0147 0.0146 (N=97)	0.0083 0.0084 (N=97)	0.0084 0.0085 (N=97)	0.0086 (N=97)
Average speed in secs (SD in parentheses)	N=100	10.80 (2.29)	31.71 (13.17)	41.21 (29.12)	457.75 (147.66)
	N=97	10.82 (2.32)	31.85 (13.35)	40.82 (28.89)	457.44 (149.91)

for all parameters except $\tau_{01,}$ and γ_{01} with respect to the GQ estimators. The relative efficiencies of the MSEs of all the other estimators are within reasonable limits of each other.

As in the univariate random effects case, the standard deviations of PQL estimates were the smallest (Table 5.12) of all the standard deviations of estimates across replicates for all parameters and methods though not by that much and the standard deviations of estimates of all the other methods didn't seem to differ from each other. This is not surprising considering the fact that PQL's estimates of all parameters displayed significant

Table 5.12 - Standard Error Estimates for the Bivariate Model

Title	Parameter	Avg SE of Estimates	SD of Estimates	Root MSE
PQL	γ_{00}=-1.2	0.1097	0.0995	0.1493
	γ_{01}=1	0.0984	0.0926	0.1456
	γ_{10}=1	0.0592	0.0594	0.1077
	τ_{00}=1.625	0.1793	0.2138	0.4354
	τ_{01}=0.1	0.0764	0.0713	0.0860
	τ_{11}=0.25	0.0641	0.0642	0.1212
Laplace6	γ_{00}=-1.2	0.1282	0.1111	0.1105
	γ_{01}=1	0.1163	0.1035	0.1039
	γ_{10}=1	0.0740	0.0682	0.0686
	τ_{00}=1.625	0.2610	0.2906	0.2910
	τ_{01}=0.1	0.1113	0.0971	0.0970
	τ_{11}=0.25	0.0919	0.0912	0.0911
Gauss	γ_{00}=-1.2	0.1197	0.1210	0.1204
	γ_{01}=1	0.1098	0.1114	0.1109
	γ_{10}=1	0.0733	0.0697	0.0693
	τ_{00}=1.625	0.2467	0.3110	0.3114
	τ_{01}=0.1	0.1101	0.1012	0.1010
	τ_{11}=0.25	0.0908	0.0912	0.0917
AGQ (N=97)	γ_{00}=-1.2	0.1260	0.1129	0.1122
	γ_{01}=1	0.1124	0.1063	0.1058
	γ_{10}=1	0.0724	0.0695	0.0693
	τ_{00}=1.625	0.2550	0.3047	0.3041
	τ_{01}=0.1	0.1068	0.1018	0.1015
	τ_{11}=0.25	0.0859	0.0924	0.0927

biases which is borne out by the fairly large discrepancies between the standard deviations and corresponding root MSEs for PQL.

The average standard errors of PQL estimates were again the smallest while the average standard errors of estimates from other methods didn't differ that much from each other. This again is due to the bias that PQL suffers in contrast to the other methods as there were no tangible discrepancies between the average standard errors and the standard deviations of estimates for all methods. The standard errors of the variance-covariance components for GQ were computed using the multivariate delta method as MIXOR gives

the standard errors of the elements of the Cholesky matrix of the variance-covariance estimates matrix.

Finally, PQL was by far the fastest method (Table 5.11) – about three times as fast as Laplace6, about four times as fast as GQ (Gauss), and a whopping 42 times as fast as AGQ – for this group of datasets, followed by Laplace6, GQ and AGQ, in that order. AGQ was by far the slowest. Besides, it failed to converge for 3 of the 100 datasets with the error message "No valid parameter points were found." It might have converged if I specified more accurate (i.e., closer to the true parameters' values) initial values for the parameters.

In summary, for this group of datasets simulated with a hierarchical logistic model with bivariate random effects, Laplace6 (as implemented in HLM) seems to perform the best overall. PQL, though by far the fastest, suffers from significant (negative) biases, and AGQ (as implemented by SAS) suffers from slowness and the need to specify accurate initial values for the parameters which can be problematic if we don't have a good hunch of what they might be. The non-adaptive Gauss is quite comparable to Laplace6 but, as implemented in MIXOR, it has the inconvenience of not directly giving the standard errors of the random effects variance-covariance components; it gives the standard error of their Cholesky ("square root") components instead. Once this is taken care of, it appears to be a good candidate to compete with Laplace6 for the case of these bivariate random effects data.

Chapter 6

AN APPLICATION IN EDUCATION

Introduction

In this chapter, real-life educational data is analyzed using the four methods -- namely, PQL, Laplace6, non-adaptive and adaptive Gauss-Hermite quadratures -- to see how they perform. The data set used is the 1988 Thailand National Survey of Primary Education (henceforth called the Thailand data). A dichotomous outcome is selected to illustrate the use of hierarchical logistic model that is the focus of this thesis.

Description of the Thailand Data

The Thailand data (USAID contract DPE-5824-A00-5076-00) is a national survey of more than 400 primary schools in Thailand conducted in 1988 by a research team from the College of Education at Michigan State University and Office of the National Educational System of the Royal Thai Government. It was "a multipurpose survey of conditions, practices, outcomes, and costs of primary schooling." (Raudenbush et al., 1993).

The survey used a multistage cluster sampling scheme. Thailand has 12 multi-province educational regions plus the Bangkok metropolitan area as the 13^{th} region. While the latter is a single province, a typical educational region includes about seven or eight provinces. Altogether, there are 72 provinces in Thailand, each one within one and only one educational region. There are a number of districts within each province, and a number of schools within each district.

The first stage of sampling involved a stratified random sample of 25% (i.e., n=18) of the provinces within strata comprising of educational regions. Twenty percent of the districts within provinces, and 30% of the schools within districts were sampled randomly. One sixth-grade class was selected at random from each sampled school (though many schools contain only one sixth-grade class) and all students within each selected class were administered a student survey questionnaire. (At the person level, samples were also drawn from three other populations: principals, teachers, and parents. Here, we are only interested in the student data.) School level data (for the schools from which the classes were drawn) was also collected. Altogether, more than 400 schools were randomly selected and data collected on almost 10,000 (sixth-grade) students within the schools (Raudenbush et al., 1993; Raudenbush and Bhumirat, 1992). Unfortunately, due to insufficient data at both school and student levels, there remained 392 schools with 8194 students for this analysis.

Since 1980 Thailand had launched programs to improve the quality of education. This included a pre-primary education program to improve student readiness for school, staff development programs for teachers, and national testing programs to hold educators accountable for student learning by requiring students to demonstrate basic skills before advancing to the next grade.

The medium of instruction in class is the central Thai dialect. Students who don't speak central Thai at home maybe disadvantaged in their schooling, an argument not unlike the advocacy for the use of ebonics to teach underprivileged African-American students. So, my research question here is whether students who don't speak central Thai

at home are more likely to repeat a grade in primary school after accounting for other relevant student and contextual (school) factors.

Formulation of Hypothesized Model

As mentioned above, it may well be that the pre-primary education program the government launched was effective. Hence, taking pre-primary education may result in less probability of repeating a grade for a student. Another relevant variable, which almost always has a positive effect on student achievement and thus might negatively affect the probability of repeating a grade, is the student's (family's, to be precise) socioeconomic status (SES). Student nutrition, as represented by whether the student has breakfast everyday, might have an effect on the student's attention in class and hence on his performance. Traveling time from home to school is another interesting variable to consider since students whose homes are far away from school, especially the ones living in rural areas, are more likely to come to school late or be absent from school. Finally, it would be interesting to see if there are gender differences in the probability of repeating a grade.

In summary, I will use the following the student-level variables in my model:

outcome variable – whether student repeated grade(s) in primary school (REP, 1=yes, 0=no);

research variable – whether the student speaks central Thai at home or not as reported by student(DIALECT, 1=central Thai, 0=other);

covariates – pre-primary experience (PPED, 1= 1 or more years of pre-primary education, 0=none) based on student report; SES (SESC, derived from measures of parents'

education and occupation as well as the natural logarithm of the amount of pocket money the student typically brings to school as reported by student, grand mean centered); nutrition (BRF, 1=student eats breakfast daily, 0=not daily); time needed to travel from home to school (in hours) (L_HSTC, log and centered); and gender (DSSEX, 1=male, 0=female). Thus, my hypothesized model at level-1 (student level) would be

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_{0j} + \beta_{1j}(SESC)_{ij} + \beta_{2j}(L\_HSTC)_{ij} + \beta_{3j}(DSSEX)_{ij} + \beta_{4j}(DIALECT)_{ij} + \beta_{5j}(BRF)_{ij} + \beta_{6j}(PPED)_{ij} \quad \textbf{(6.1)}$$

Student learning is not only influenced by student factors but also by relevant environmental factors. Some school factors that would contribute to student learning include availability of facilities in school, availability of textbooks and mean school SES. The availability of facilities in school (ZFACTOTC) is an aggregated variable derived from 18-item scale (and grand-mean centered) including primarily equipment used for instruction ("hard technologies") but also some equipment that could be used for administration. The items included the presence or absence of a Thai typewriter, English typewriter, copying machine, slide projector, overhead projector, amplifier, radio cassette, radio, tape module, television, etc. The availability of textbooks and workbooks (MTXTBKC) is the sum of the texts and workbooks available for student use (as reported by student) across the five areas of the curriculum (and averaged for school) and then grand mean centered. Student SES was aggregated to the school level and then grand mean centered to create mean school SES (MSESC).

Note that the school level variables are for grade 6, i.e., they were taken (measured) when the student was in grade 6. However, the outcome variable pertains for the whole duration of the student's primary education. Therefore, the assumption is made here that there isn't much mobility of students across schools during the students' primary school years.

Besides these three school level variables directly (i.e., through the level-1 intercept) having an effect on the log-odds of repetition for a student, I am hypothesizing that some of the level-1 effects may be in some way affected by one or more of these level-2 predictors. The student's SES effect on the log-odds of grade repetition may be mediated by the student's school mean SES, in the sense that the effect of SES on the log-odds of repetition for a low-SES student attending a high mean SES school may be offset by the advantage of attending the high mean SES school. The availability of facilities and textbooks might likewise affect the effect of the student's SES on the student's log-odds of repeating a grade. The latter two school-level variables are also hypothesized to mediate the effect of dialect, the research variable, on the log-odds of repetition. Furthermore, only the level-1 intercept term is hypothesized to have a random effect at level 2. This decision was made after a preliminary run where random effects terms that were added at level 2 for the coefficients of the only two non-dichotomous covariates at level-1, namely, SESC and L_HSTC, turned out to be non-significant. So, my hypothesized model at level-2 (school level) would be

$$
\begin{aligned}
\beta_{0j} &= \gamma_{00} + \gamma_{01}(MSESC)_j + \gamma_{02}(ZFACTOTC)_j + \gamma_{03}(MTXTBKC)_j + u_{0j}, \quad u_{0j} \sim N(0,\tau) \\
\beta_{1j} &= \gamma_{10} + \gamma_{11}(MSESC)_j + \gamma_{12}(ZFACTOTC)_j + \gamma_{13}(MTXTBKC)_j \\
\beta_{2j} &= \gamma_{20} \\
\beta_{3j} &= \gamma_{30} \\
\beta_{4j} &= \gamma_{40} + \gamma_{41}(ZFACTOTC)_j + \gamma_{42}(MTXTBKC)_j \\
\beta_{5j} &= \gamma_{50} \\
\beta_{6j} &= \gamma_{60}
\end{aligned}
\tag{6.2}
$$

Results

The descriptive statistics of the variables used in this analysis are summarized in Table 6.1. My hypothesized model was estimated using the four methods and the estimates

Table 6.1 - Descriptive Statistics for the Thailand Data

Student Level					
VARIABLE	N	MEAN	SD	MINIMUM	MAXIMUM
SESC	8194	-0.00	0.69	-1.76	3.48
L_HSTC	8194	-0.00	0.71	-2.74	1.97
DSSEX	8194	0.51	0.50	0.00	1.00
DIALECT	8194	0.47	0.50	0.00	1.00
BRF	8194	0.84	0.37	0.00	1.00
PPED	8194	0.49	0.50	0.00	1.00
REP	8194	0.14	0.35	0.00	1.00
School Level					
VARIABLE	N	MEAN	SD	MINIMUM	MAXIMUM
ZFACTOTC	392	0.00	0.39	-0.88	1.50
MSESC	392	-0.00	0.45	-0.93	2.01
MTXTBKC	392	-0.01	1.82	-5.91	2.63

of the parameters (fixed effects as well as random effects variance) and their standard errors from the four methods are summarized in Table 6.2.

Table 6.2 - Estimates for the Hypothesized Model

Parameter	PQL	Laplace6	Gauss-10	AGH
Intercept	-2.036* (0.137)	-2.223* (0.138)	-2.256* (0.141)	-2.218* (0.147)
MSESC	-0.989* (0.214)	-1.176* (0.249)	-1.155* (0.260)	-1.094* (0.231)
ZFACTOTC	0.295 (0.243)	0.333 (0.251)	0.286 (0.253)	0.303 (0.263)
MTXTBKC	0.073 (0.051)	0.086 (0.055)	0.076 (0.055)	0.079 (0.055)
SESC	-0.566* (0.100)	-0.544* (0.107)	-0.602* (0.107)	-0.592* (0.103)
MSESC	0.388* (0.161)	0.406* (0.186)	0.438* (0.188)	0.413* (0.168)
ZFACTOTC	0.546* (0.208)	0.772* (0.277)	0.540* (0.274)	0.572* (0.214)
MTXTBKC	-0.021 (0.054)	-0.018 (0.046)	-0.010 (0.047)	-0.023 (0.056)
L_HSTC	0.083 (0.054)	0.103 (0.063)	0.083 (0.063)	0.089 (0.056)
DSSEX	0.588* (0.071)	0.597* (0.070)	0.619* (0.071)	0.616* (0.073)
DIALECT	0.206 (0.122)	0.239 (0.129)	0.270* (0.125)	0.234 (0.130)
ZFACTOTC	-0.082 (0.305)	0.005 (0.326)	0.033 (0.327)	-0.085 (0.325)
MTXTBKC	-0.161* (0.071)	-0.191* (0.086)	-0.214* (0.082)	-0.172* (0.075)
BRF	-0.387* (0.099)	-0.394* (0.099)	-0.392* (0.101)	-0.403* (0.102)
PPED	-0.383* (0.092)	-0.418* (0.096)	-0.423* (0.096)	-0.411* (0.096)
ran.eff.variance	1.038* (0.113)	1.351* (0.167)	1.425* (0.162)	1.290* (0.159)

* Significant at the 5% level. Standard errors are shown in parentheses.

My research hypothesis of whether using the Thai dialect at home or not has an effect, adjusting for other covariates, on the log-odds of repeating a grade has ambiguous results. While it is significant, at the 5% level, using Gauss with 10 quadrature points (p-

value = .032), it didn't turn out to be significant for the other methods, though the p-values were close to 0.05 (0.09 for PQL, 0.063 for Laplace6, and 0.072 for AGQ). So, maybe the effect of dialect on the log-odds of repeating a grade may be considered marginally significant, i.e., there is an *inconclusive* evidence that speaking central Thai at home increases the log-odds of repeating a grade for a student. I suspect this might have been a fluke for when I reran Gauss with 20 quadrature points, the effect of dialect on the log-odds of repetition turned out to be non-significant (p-value=0.069). Even so, textbooks has a significant negative effect on the log-odds of repetition indicating that not speaking central Thai at home can be offset by the availability of textbooks. Note that the availability of facilities in school didn't have a significant effect on the log-odds of repetition.

Higher mean school SES reduces the log-odds of repeating a grade while neither the availability of facilities nor textbooks has an effect on it. The student's SES also reduces the student's log-odds of repetition and its effect (the coefficient) is positively affected by both mean school SES and the availability of facilities in school but not by textbooks.

The time it takes the student to go from home to school (given in log-hours) didn't seem to have an effect on the odds of the student's repetition. On the other hand, the gender of the student had a positive effect on the log-odds of repetition. That is, male students are more likely to repeat a grade (perhaps due to the fact they are more likely to play truant and less serious about school.) As expected eating breakfast regularly reduces

the log-odds of repeating a grade for a student. And, so does pre-primary educational experience by the student.

Finally, the random effects variance was found to be significant by all the methods indicating unaccounted random variation among the schools in the level-1 intercept for the log-odds of student repetition.

In regards to the comparison of methods as far as estimation is concerned, there didn't appear to be that much difference among the methods in the fixed effects estimation. However, the random effects variance was severely underestimated by PQL as compared to the other methods. Incidentally, the SAS algorithm that implemented the adaptive Gauss adaptively selected one quadrature point to run the adaptive Gauss-Hermite. The results didn't change appreciably when I forced AGH to run with five quadrature points.

Chapter 7

DISCUSSION AND CONCLUSION

This dissertation compared four methods, two Laplace-based and two Gaussian-based, for approximating the likelihood for multilevel logistic models (mixed logistic models). The comparison was done graphically, analytically (i.e., in terms of asymptotic properties), as well as via simulation studies. The methods were also applied on real educational data to see what kinds of results they gave and how close to each other they are.

First, an illustrative example of a very simple mixed logistic model was given and the likelihood derived but not in closed form. Since the integral involved in the likelihood cannot be done in closed form -- this is what the dissertation attempted to address -- how each method approximates this likelihood was treated in a fairly detailed manner. In fact, this was also done for two more Laplace-based methods. The various integrand approximations to the integrand in the likelihood were derived for the various methods and plotted against the actual integrand in the likelihood. The numerical integral approximations to the "true" integral (obtained by Trapezoidal Rule with error $< 10^{-8}$) were also computed for the various methods.

For the Laplace-based methods, the integrand approximation functions (of the random effect) of order 2, 4, 6 and 8 were derived and plotted along with the actual integrand to show how each progression closely approximates the integrand. The plot showed that even the graph of the second-order Laplace was already close enough to the actual integrand. The numerical integral approximation got better by the order of 10 as the

order of the Laplace increased up to 6 and then stabilized. In fact, the error of Laplace6 was smaller than that of Laplace8. Of course, this is just one example. Besides, the approximations were stopped at order 8 and higher orders were not investigated.

For the Gaussian-based methods, the general integrand approximation functions were first derived for both the non-adaptive and adaptive Gauss-Hermite methods. The non-adaptive Gauss-Hermite integrand approximation formula was plotted for numbers of quadrature points starting from 2 along with the actual integrand function. The successive graphs showed it takes quite a number of quadrature points (more than 10) for the non-adaptive Gaussian integrand to get really close to the actual integrand. This was borne out by the numerical integral computation which took about 8 quadrature points to have an (absolute) error similar to that of Laplace2 and about 14 quadrature points to get an error close to that of Laplace6. The adaptive Gaussian approach first transforms (standardizes) the integrand itself to one centered around zero. This transformed integrand, which I labeled the AGH integrand, is the one that the adaptive Gauss integrand approximations try to get close to as the number of quadrature points increases. The graphs of the AGH integrand along with the integrand approximations display that the integrand approximations are already quite close to the AGH integrand for 2 quadrature points. The numerical integral computation shows that 4 quadrature points were enough to get an error of the same order as Laplace6. (Of course, adaptive Gauss with one quadrature point gives an identical value as Laplace2.)

Next, the general mixed logistic model was formulated and how each of the four methods approximate the likelihood described. Then, for the single random effect case, the

asymptotic behaviors of the three methods that are based on centering the random effect around its conditional mode, namely Laplace2, Laplace6 and adaptive Gauss-Hermite, were worked out. It turned out that the errors of the three methods for each cluster (since the integration is approximated for each cluster by the methods) were $O(n^{-1})$, $O(n^{-2})$ and $O(n^{[-G/3]})$ respectively where n is the cluster size, G is the number of quadrature points and [] is the greatest integer function. This implies that the errors of Laplace2 and adaptive Gauss are of the same order when G=1 and those of Laplace6 and adaptive Gauss when G=4 which seems to confirm the example in the previous chapter. In fact, the adaptive Gauss-Hermite with one quadrature point and Laplace2 are identical, i.e., have the same formula.

A simulation study was done for both a univariate random effect hierarchical logistic model and a bivariate random effects one. First, eight groups of 100 datasets, each group representing one of two values (small and large) of random effects variance, conditional probability and cluster size were simulated using a univariate random effect model. Algorithms implementing the four methods were then run on each dataset and their performance was investigated. All methods performed quite well when the cluster size was large except that PQL (the way Laplace2 was implemented in HLM) was usually the most biased and sometimes had the largest mean-squared error. In this case, the non-adaptive Gauss was the fastest, followed by Laplace2, Laplace6 and adaptive Gauss, in that order. I think this was due to an implementation fluke. After all, SAS (which was used for the adaptive Gauss) is a general purpose statistical software while MIXOR (implementing the non-adaptive Gauss) is a specialized program. Had I used the SAS non-linear mixed

program with the non-adaptive option, it would most likely have taken more time than the default adaptive option since it would need more quadrature points. (I had chosen the non-adaptive option on one dataset before and then resorted to the default adaptive because it took more time.) The speeds of Laplace2 and Laplace6 make sense since the initial values of Laplace6 are those of the Laplace2 estimates (i.e., end results.)

When the cluster size was small, the adaptive Gauss gave the best results overall being the fastest among all the methods, having the least mean-squared errors and being no more biased than the other methods. In this case, the non-adaptive Gauss was the worst, not even giving results on many occasions (especially when the random effects variance was small). When it gave results in these instances, the estimates were sometimes biased, and when both random effects variance and the conditional probability of success were small, the estimates were always biased.

Another group of 100 datasets was simulated using a bivariate random effects hierarchical logistic model and programs implementing the four methods were run on this group. In this case, Laplace6 appeared to perform the best overall, especially in terms of bias and MSE. The non-adaptive Gauss was quite close to it, too. Predictably, PQL was by far the fastest for this group; however, it had the most biases (negative and significant). Adaptive Gaussian was the slowest and suffers from the need to specify good initial values for the parameters (at least as implemented in SAS).

Finally, the algorithms that implemented the four methods were run on real-life data, the 1988 Thai National Survey of Primary education to estimate a hypothesized hierarchical logistic model with a single random effect. The results indicate that the four

methods gave similar results in terms of significance, i.e., whatever was significant in under one method was also significant under the others. The Laplace2 random effects variance estimate was pretty much smaller than its counterparts from the other methods which were quite close to each other.

As far as speed of methods is concerned, it is worth noting that, unlike the graphical and analytic comparisons, the programs that were used in the simulation study and data analysis involved maximization of the likelihood as well. The maximization procedures used by the programs differ from each other and, as these procedures are usually iterative, they might have an effect on the speed of the "method" being compared. In other words, while dealing with the programs, I am not just comparing the speeds of the integration approximation methods but also those of the procedures used to maximize the resultant approximate integrals.

Suggestions for Future Study

The simulation study done in this dissertation suffers from at least one problem - the use of different programs for different methods. What I actually compared in the simulation study may not just be the methods but also the programs, i.e., implementation may matter. For instance, it would be more interesting and fair to compare the two Gaussian methods using the same program (e.g., SAS PROC NLMIXED with adaptive as well as non-adaptive options). One may also be interested in comparing different programs implementing the same method. For example, Pinheiro et al. (1993) had implemented their adaptive Gaussian and other procedures in S-Plus which they contributed to statlib. One might want to compare their implementation with the SAS implementation. One might

also want to compare the SAS procedure NLMIXED with the non-adaptive option with MIXOR, specifying in both cases the same numbers of quadrature points.

Of course, one might want to extend the comparisons to other methods such as Monte Carlo methods. For instance, Pinheiro et al. (1993) had included importance sampling in their implementation.

The more interesting and useful undertakings to take would be to improve some of the methods suggested in this dissertation. For instance, one might increase the terms in the Laplace-based approaches (i.e., increase the order of Laplace) and see where these approaches stabilize, if at all. Since the additional terms are difficult to derive, it would be useful to see if adding terms would be worth the effort in terms of improving accuracy.

In this connection, Liu and Pierce (1994) conjectured that the *m*-order (adaptive) Gauss-Hermite quadrature can be thought of alternatively as the form of '*m*-order Laplace approximation'. It would be worthwhile to investigate this assertion because, if true, the *m*-order adaptive Gauss-Hermite can be substituted for the *m*-order Laplace which the authors suggest is more preferable in applied work. This would be desirable because the latter involves the derivation of additional cumbersome and fairly difficult terms as the order increases. This assumes that the *m*-order Laplace is done the same way as that of Yang (1998). It would also be interesting to see if one can come up with another way to directly define and estimate the '*m*-order' Laplace.

This dissertation dealt only with two-level hierarchical logistic models. None of the methods (to be precise, the algorithms implementing them) except PQL (Laplace2) has implemented the three-level counterpart. It would be a good research topic to derive the

formulas (and algorithms) needed for any of the other three methods to estimate the three-level hierarchical logistic model.

Finally, one can expand the methods (and the programs) to handle generalized linear mixed models (GLMMs) rather than just hierarchical logistic models. I am aware that Laplace2 and adaptive Gauss-Hermite (really, HLM and SAS) can handle GLMMs. By default, since SAS has a non-adaptive option for its Gaussian procedure, the non-adaptive Gauss can also be implemented for GLMMs. So, one is left with Laplace6 (and higher-order Laplaces) to work on to handle GLMMs. Raudenbush et al. (2000) indicate this might not be difficult for count data.

REFERENCES

Aitkin, M., Anderson, d. and Hinde J. (1981). Statistical modeling of data on teaching styles (with discussion). *Journal of the Royal Statistical Society*, A **144**, 148-61.

Anderson, D. and Aitkin, M. (1985). Variance component models with binary response: Interviewer variability. *Journal of the Royal Statistical Society*, B **47**, 203-210.

Bennet, N. (1976). *Teaching Styles and Pupil Progress*. London: Open Books.

Bock, R.D. (Ed.) (1989). *Multilevel Statistical Methods in Educational Research*. New York: Academic Press.

Bock, R.D., Gibbons, R.D., and Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, **12**, 261-80.

Breslow, N.E., and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 421, 9-25.

Breslow, N.E. and Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrics*, **82**, 81-91.

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

Bryk, A.S., Raudenbush, S.W., and Congdon, R. (2000). *HLM: Hierarchical Linear and Nonlinear Modeling,* Version 5.20. Chicago: Scientific Software International, Inc.

Bryk, A.S. and Thum, Y.M. (1989). The effects of high school organization on dropping out: an exploratory investigation. *American Educational Research Journal*, **26**, 353-383.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences,* 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates.

Davis, P.J., and Rabinowitz, P. (1984). *Methods of Numerical Integration*, 2nd ed. Orlando: Academic Press.

Dempster, A.P., Rubin, D.B., and Tsutakawa, R.K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association*, 76, 341-53.

Evans, M., Hastings, N., and Peacock, B. (1993). *Statistical Distributions*, (2nd ed.), New York: John Wiley & Sons, Inc.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73, 43-56.

Goldstein, H. (1987). *Multilevel Models in Educational and Social Research*. London: Oxford University Press.

Goldstein, H. (1991). Nonlinear multilevel models, with an application to discrete response data. *Biometrika*, 78, 45-51.

Goldstein, H. (1995). *Multilevel Statistical Models*. New York: Halstead.

Green, P.J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review*, 55, 245-59.

Hedeker, D., and Gibbons, R.D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics*, **50**, 933-44.

Hedeker, D., and Gibbons, R.D. (1996). MIXOR: A computer program for mixed-effects ordinal probit and logistic regression analysis. *Computer Methods and Program in Biomedicine*, **49**, 157-176.

Horney, J., Osgood, D. W., and Marshall, I.H. (1995). Criminal careers in the short-term: intra-individual variability in crime and its relation to local life circumstances. *American Sociological Review*, **60**, 655-673.

Liu, Q. (1993). Laplace approximations to likelihood functions for generalized linear mixed models. Unpublished dissertation paper: Oregon state University.

Liu, Q., and Pierce, D.A. (1994). A note on Gauss-Hermite quadrature. *Biometrika*, 81, 624-629.

Longford, N.T. (1993). *Random Coefficient Models*. Oxford: Clarendon Press.

Longford, N.T. (1994). Logistic regression with random coefficients. *Journal of Computational Statistics and Data Analysis*, **17**, 1-15.

Louis, K.S., Marks, H.M., and Kruse, S. (1994). Teachers' professional community in restructuring schools. Center on Organization and Restructuring of Schools. (ERIC Document Reproduction Service No. ED 381 871)

Mathews, J.H. (1987). *Numerical Methods for Computer Science, Engineering, and Mathematics*. Englewood Cliffs: Prentice-Hall.

McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models* (2nd Ed.), London: Chapman and Hall.

McCulloch, C.E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92, 162-170.

Pinheiro, J.C., Bates, D.M., and Lindstrom, M.J. (1993). Nonlinear mixed effects classes and methods for S. *Technical Report No. 906*, University of Wisconsin-Madison, Dept. of Statistics.

Pinheiro, J.C., and Bates, D.M. (1995). Approximations to the log-likelihood function in the nonlinear mixed effects model. *Journal of Computational and Graphical Statistics*, **4**(1), 12-35.

Raudenbush, S.W., and Bryk, A.S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, 59, 1-17.

Raudenbush, S.W., and Willms, J.D. (1991). *Pupils, Classrooms, and Schools: International Studies of Schooling from a Multilevel Perspective*. New York: Academic Press.

Raudenbush, S., and Bhumirat, C. (1992). The distribution of resources for primary education and its consequences for educational achievement in Thailand. *International Journal of Educational Research*, *17*(2), 143-164.

Raudenbush, S.W.(1993). Posterior modal estimation for hierarchical generalized linear models with application to dichotomous and count data. Unpublished manuscript.

Raudenbush, S.W., Eamsukkawat, S., Di-ibor, I., Kamali, M., and Taoklam, W. (1993). On-the-job improvement in teacher competence: Policy options and their effects on teaching and learning in Thailand. *Educational Evaluation and Policy Analysis*, 15(3), 279-297.

Raudenbush, S.W. (1999). Hierarchical models. In S. Kotz (Ed.), *Encyclopedia of Statistical Sciences, Update Volume 3* (pp. 318-323). New York: John Wiley & Sons, Inc.

Raudenbush, S.W., Yang, M., and Yosef, M. (2000). Maximum likelihood for hierarchical models via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics*, **9**(1), 141-157.

Rosenberg, B. (1973). Linear regression with randomly dispersed parameters. *Biometrika*, 60, 65-72.

Rodriguez, G., and Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of Royal Statistical Society*, A, **158** (1), 73-89.

Rumberger, R.W. (1995). Dropping out of middle school: a multilevel analysis of students and schools. *American Educational Research Journal*, **32**, 583-625.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78, 719-727.

Scheid, F. (1968). *Theory and Problems of Numerical Analysis*. New York: McGraw-Hill.

Stiratelli, R., Laird, N., and Ware, J.H. (1984). Random effects models for serial observations with binary response. *Biometrics*, **40**, 961-71.

Stroud, A.H., and Secrest, D. (1966). *Gaussian quadrature formulas*. New Jersey: Prentice-Hall.

Tierney, L., and Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of American Statistical Association*, 81, 82-86.

Waldman, D.A., and Avolio, B.J. (1991). Race effects in performance evaluations: controlling for ability, education and experience. *Journal of Applied Psychology*, 76, 897-901.

Wolfinger, R. (1999). Nonlinear mixed models: a future direction. Presented at the Annual Interface Conference, Shaumberg, IL, June 10, 1999.

Yang, M. (1998). *Increasing the efficiency in estimating multilevel Bernoulli models*. Unpublished dissertation paper, College of Education, Michigan State University.

Yosef, M. (1997). *Two-level hierarchical mixed-effects logistic regression analysis: a comparison of maximum likelihood and penalized quasi-likelihood estimates*. Unpublished apprenticeship paper, College of Education, Michigan State University.

Young, D.J. (1996). Science achievement and educational productivity: a hierarchical linear model. *Journal of Educational Research*, 8, 272-278.

Zeger, S.L., Liang, K. and Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, **44**, 1049-60.