This is to certify that the

dissertation entitled

Increasing the Efficiency in Estimating Multilevel Bernoulli Models

presented by

Meng-Li Yang

has been accepted towards fulfillment of the requirements for

Ph. D. degree in Measurement and Quantitative Methods

Stephen W. Raudenbush
Major professor

Date May 15, 1998

INCREASING THE EFFICIENCY IN ESTIMATING MULTILEVEL BERNOULLI MODELS

By

Meng-Li Yang

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

Spring 1998

ABSTRACT

INCREASING THE EFFICIENCY IN ESTIMATING
MULTILEVEL BERNOULLI MODELS

by

Meng-Li Yang

Multi-level linear models are useful tools for educational research, where observations are often nested within clusters. If both the response and the random effects have normal distributions, maximum likelihood inferences for the fixed and random effects variances can be obtained analytically. For dichotomous responses such as dropping out and repeating a grade, logistic regression is often used to model the relationship between the responses and the covariates. However, in such multilevel Bernoulli models, estimation has been a problem. Since the responses have a Bernoulli distribution, which is not conjugate to the normal distribution of the random effects, rough approximation or numerical integration has to be used to approximate the marginal distribution of the responses in order to obtain maximum likelihood estimates. The strategies proposed before include the penalized quasi-likelihood approach, the approximate maximum likelihood approach using Monte Carlo methods or the Gauss-Hermite Quadrature technique, and the Bayes approach.

This dissertation proposes using Laplace approximation to the marginal distribution and then using approximate Fisher scoring to find the maximum likelihood inferences for the parameters. To achieve the goal, first, the infinite multivariate Taylor series is deduced. Via the infinite multivariate Taylor series, Laplace approximation can be extended to any order and any dimension. However, through preliminary experiments, approximation up to the sixth order is found to produce sufficiently accurate estimates. The resultant program is therefore called Laplace6.

Laplace6 is investigated using various simulated data sets by comparing its estimates with those of PQL, PQL2 and Gauss-Hermite Quadrature. Laplace6 was found to have, generally, the highest efficiencies among all the methods compared. The 1988 National Survey of Primary Education in Thailand was also analyzed using all of the above programs. Laplace6 estimates were found to be close to those produced by Gauss-Hermite Quadrature using 30 and 40 quadrature points. In addition, to check the consistency property of the approximate maximum likelihood estimates produced by Laplace6, 400 bivariate data sets were generated. Half of the 400 had 200 clusters in the second level and the other half had 2000 clusters. Laplace6 estimates were found to be normally distributed with small negative bias. Moreover, the variances of the estimates of the data sets with 2000 clusters were 10 times smaller than those with 200 clusters. Thus, Laplace6 estimates were approximately consistent.

To my grandmother and my father

ACKNOWLEDGMENTS

I would like to express my deepest gratitude for my advisor, Dr. Stephen W. Raudenbush for his enormous amount of instruction, guidance, support and push throughout my doctoral program. The dissertation would not exist without him. My special thanks go warmly to Richard Congdon, our computer programmer, who is so efficient and so pleasant to work with.

I also want to thank my committee, Dr. James Stapleton, Dr. Ken Frank and Dr. Alka Indurkhya for their comments and suggestions. Dr. Stapleton's suggestions about improving the last part of the simulation study especially helped provide more useful information about the behavior of the approximation method proposed here.

I have always wanted to say "Thank you" to Dr. Shyh-Leh Chen, who is now an associate professor in Taiwan, and my colleague, Matheos Yosef. Their kindness and knowledge helped me go through the confusion and awkwardness of a freshman in applied statistics. I am indebted to my friend, Dr. Su-hao Tu for her support and friendship for the past few years.

I am grateful that my parents allowed me to spend so many years studying abroad. I don't think I can ever fully appreciate their love for me. I am thankful to God that I have wonderful sisters and brother to grow up together with through happier times and harder times.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1

INTRODUCTION

People are concerned about the quality of education. They want to know what factors --- educational policies, programs, school environments, characteristics of teachers, or instructional approaches --- contribute to students' best learning. Educational research tries to find answers to these concerns. Large samples of students from different classrooms or schools are often drawn in order to support generalizable conclusions. However, students are nested within classrooms, classrooms are nested within schools, and schools within districts. Thus student learning is embedded within clusters, i.e., classrooms or schools. Because each cluster has a special climate due to its components, such as students and teachers, not only will individual students differ from one another, but there will be group differences among clusters. Longitudinal data, with repeated measurements from the same person, can also be regarded as nested data. Here each person is considered a cluster, with observations of the same person more similar than observations from different people.

Bennett (1976) found a significant difference between two styles ('formal' and non-formal) of teaching when he used multiple regression analysis, ignoring the grouping of the students into classes. However, when Aitkin et al. (1981) analyzed the same data

accounting for the nesting effect, the difference disappeared. In fact, with such a nested data structure, traditional statistical tools such as ANOVA and multiple regression either have great limitations or cannot work appropriately. Instead, as a statistical tool developed specifically for nested designs, hierarchical or multilevel linear regression models, also referred to as random coefficient models (Rosenberg, 1973) or covariance components models (Dempster, Rubin, and Tsutakawa, 1981), allow each cluster to have its own slope and regression coefficients. These slopes and regression coefficients are often considered normally distributed with mean equal to the effects of cluster characteristics as specified in the higher level, between-cluster model. That is, in a higher level these coefficients along with the slopes are each predicted by a set of cluster characteristics. The errors from the prediction are referred to as the random effects, normally distributed with mean zero and a variance covariance matrix.

For example, in the first level of a 2-level model, observation j in the ith cluster, y_{ij}, is modeled by a vector of independent variables, x_{ij}, $y_{ij} = x_{ij}^T \alpha_i + e_{ij}$, where α_i is a vector of the regression coefficients (including the intercept) for the ith cluster; e_{ij} is the random term for y_{ij}, assumed to be normally distributed with mean 0 and variance σ^2, $e_{ij} \sim N(0, \sigma^2)$. In the second level, α_i from each cluster is collected together as the dependent variable to be predicted by cluster characteristics. Assume $\alpha_i = \begin{bmatrix} \alpha_{i1} \\ \alpha_{i2} \end{bmatrix}$. Then there will be two equations for level-2: $\alpha_{i1} = w_i^T \beta_1 + b_{1i}$ and $\alpha_{i2} = z_i^T \beta_2 + b_{2i}$, where w_i and z_i are both vectors of the ith cluster characteristics; β_1 and β_2, are vectors of

regression coefficients for α_1 and α_2, respectively; b_{1i} is the random effect of cluster *i* for α_1 and b_{2i} the random effect of cluster *i* for α_2. b_1 and b_2 are assumed to be multivariately normally distributed with mean 0 and a variance matrix *D*.

Such modeling gives educational researchers a clearer look at the mechanism of the interactions among students, teachers, schools, society and policies, and resolves the statistical difficulties encountered by ANOVA and multiple regression. The resulting variance components from each level, additionally, give information about its respective amount of unexplained variation. Such models for continuous outcomes have been well developed by researchers using different estimation methods. For example, Raudenbush (1984) used EM algorithm; Goldstein (1986) used iterative generalized least squares; deLeeuw and Kreft (1986) and Longford (1987) used Fisher scoring. A brief review reveals different applications of the models: school effectiveness as related to student achievement scores (Raudenbush and Bryk, 1986; Aitkin and Longford, 1986; Young, 1996), school effects and their stability (Raudenbush and Willms, 1991), how teacher interaction outside the classroom affects student learning (Louis, 1994), program evaluation (Marks, 1995; Lee,1995; Mac Iver and Plank, 1996), adolescent attitude change toward deviance (Raudenbush and Chan, 1993), and the effects of ratee and rater race on performance evaluations (Waldman and Avolio, 1991). Goldstein (1987), Bryk and Raudenbush (1992) and Longford (1993) gave detailed accounts of applications and methodology of these models in social and educational contexts. Bock (1989) and Raudenbush and Willms (1991) provided applications in education.

Nevertheless, although it is easy to conceive a continuous and normal distribution for human characteristics, such as intelligence, abilities and achievements, certain types of valuable educational data cannot be normally distributed. For example, whether a student has repeated a grade, dropped out of school, been admitted to college or persisted in the pursuit of higher education, might all be informative about the educational environment and policies. Since these outcomes are either yes or no (typically coded as 1 or 0), the usual linear model that assumes a normal random error fails. Instead, the logistic regression, one of the generalized linear models (McCullagh and Nelder, 1989) is used to model such outcomes. The logistic model uses the logit (the log of the odds ratio) of the dependent variable as the outcome variable. For example, to model the probability of dropping out for student j in school i using his personal information, x_{ij}, the model will be $\log(\frac{\mu_{ij}}{1-\mu_{ij}}) = x_{ij}^T \alpha_i + e_{ij}$, where $\mu_{ij} = E(y_{ij} = 1|b_i)$ is the conditional probability of dropping out, y_{ij} being the observed data with dropout = 1, non-dropout = 0.

For analyzing nested non-normal data such as binary data and count data, the multilevel generalized linear model with random effects is a natural outgrowth of both generalized linear models and hierarchical linear models. It incorporates generalized linear models into the framework of the hierarchical linear model. In the first level, the linear regression model is substituted by a generalized linear model while the second level remains a linear model. For such models, researchers face a major task of obtaining a good estimate of the marginal distribution of the data. This marginal distribution is the

integral of the product of the first level likelihood, $f(y_i|b_i)$, and the assumed second level distribution, $p(b_i)$ with respect to the random cluster effects, i.e., $\int f(y_i|b_i)p(b_i)db_i$. In most cases, the second-level model is assumed to have a normal distribution. The difficulty in evaluating the integral arises from the fact that the normal distribution is not the conjugate prior for non-normal distributions such as binomial and Poisson that are assumed in generalized linear models.

Many statisticians have studied estimation in generalized linear models with random effects, especially in the logistic model with random effects (e.g., Stratelli, Laird, and Ware, 1984; Wong and Mason, 1985; Schall, 1991). What makes the logistic model with random effects interesting and difficult is that there is no closed form for the marginal distribution of the outcome for a logistic model (Zeger et al, 1988). As a result, the estimation of the parameters, including the variance components and the fixed effects, have to be derived through approximation, if not through intensive Monte Carlo computation. Moreover, because the approximations are generally rough, the resulting estimates of the variance components is often subject to underestimation (Breslow and Clayton, 1993; Rodriguez and Goldman, 1995), thus resulting also in the underestimation of the fixed effects coefficients, especially when the number of random effects increases with the sample size and the binomial denominator is small (Breslow and Lin, 1995).

Breslow and Lin (1995) proposed first-order and second-order Laplace approximations of the integral of the marginal likelihood with in the context of asymptotic bias correction, with the second-order being the fourth order Taylor expansion

whereas the first-order being the second order Taylor expansion. Incidentally, the first-order approximation has terms exactly the same as those in the penalized quasi-likelihood approximation (Breslow and Clayton, 1993). Therefore, the second order approximation can be regarded an improvement over the penalized quasi-likelihood approach such as Raudenbush's posterior modal estimation method (1993). Later Lin and Breslow (1996) extended the approximation to the case of multiple random effects, with zero correlation among them, however. Nevertheless, the assumption of a zero correlation among random effects limits the use of the approximation in real world research.

This dissertation will build on the work of Breslow and Lin (1995) and Lin and Breslow (1996), using Magnus's (1988) and Magnus and Neudecker's ideas (1988) as the toolbox. It will generalize the Laplace approximation to multiple random effects with a general variance-covariance matrix. Moreover, through simple simulations it was found that the contribution of the eighth order term in the eighth-order expansion to the approximate log- likelihood is negligible while those of the lower orders are not. Therefore, this dissertation will also expand up to the sixth order of the Taylor series to get a satisfactory Laplace approximation to the log-likelihood. However, the purpose of the extension will not be bias correction. Treating the resulting approximate marginal likelihood as the exact likelihood, Fisher scoring will be applied to simultaneously estimate the fixed effects and the variance-covariance matrix of the random effects. Because the Laplace approximation is not stable in some cases, according to Breslow and Lin (1995), to ensure convergence and more efficient estimation, the output from Raudenbush's posterior modal algorithm (1995) will be used as starting values for

parameters to be estimated. The resulting estimation method will be called "Laplace6". It will be tested and evaluated with extensive simulation studies. Its efficiency, in terms of mean squared errors, will also be compared with those of Raudenbush's posterior modal estimation (1993), which is equivalent to the PQL (PQL) (Breslow and Clayton, 1993), of Goldstein and Rasbash's (1996) second-order penalized quasi-likelihood (PQL2) method, and approximate maximum likelihood method using Gaussian Quadrature technique (Gauss) by Hedeker and Gibbons's MIXOR(1994, 1996)).

To achieve the above goals the dissertation will

- derive the multivariate Taylor series;
- derive the six moments of the multivariate normal distribution;
- prove that the fourth and sixth moments can be substituted by simpler forms for use in the approximation;
- find the approximate log-likelihood using a Laplace sixth-order expansion of the joint density of the data and the random effects;
- find first derivatives of the approximate log-likelihood for both the fixed effects and random effects variance matrix in order to use approximate Fisher scoring to estimate the fixed effects and the random effects variance matrix;
- work out a computational algorithm, based on the derivatives and the approximate log-likelihood, for computer programming;
- analyze the data set of 1988 National Survey of Primary Education in Thailand (Thailand data) using the above methods as an example;

- generate data sets with different models and parameters, the structure of which generally follows that of Rodriguez and Goldman (1995); and
- investigate the performance of the methods by analyzing the estimators in terms of their biases and mean squared errors.

According to experience so far, when the variance and especially the conditional expectation, $E(y_{ij}|b_i)$, are both very small(e.g., .25, .01, respectively), all of the above methods except for PQL have difficulty converging. On the other hand, data sets with extremely small random effect variance and conditional expectation will not be of much practical as well as theoretical interest, anyway. Moreover, because of the symmetry of probability in dichotomous situations, a conditional expectation higher than .5 is the same as itself minus .5. Therefore, I will limit the range of the conditional expectation to (.1, .3) and the variance to (.25, 2). In addition, for a single variance component model, the within-group sample size will be around 20 and between group sample size around 150. For a model with variance-covariance matrix, even larger within- and between- group sample sizes are necessary.

Chapter 2

BACKGROUND AND SIGNIFICANCE

Among the members of the generalized linear models with random effects, the logistic model with random effects especially poses numerical difficulties. The obstacle occurs when the marginal likelihood is needed for estimation of the parameters. The marginal likelihood is obtained by integrating out the random effects from the joint likelihood of the data and the random effects. In the logistic model with random effects, the data have a Bernoulli distribution, while the random effects are usually assumed to have a multivariate normal distribution. Besides, while the conditional expectation of the response, $\mu_{ij} = E(y_{ij} = 1|b_i)$, and the sum of the fixed and random effects are linked by a canonical link function (McCullagh and Nelder, 1989) for each member, the marginal expectation, $E(y_{ij})$, is not. Researchers cannot find an exact closed form relationship between the logit link and the marginal expectation (Zeger et al., 1988). Hence there have been different approaches for estimation. A brief review with reference to the various approaches highlights the difficulty.

Bayes Approach

Zeger and Karim (1991) used the Bayesian paradigm by applying Gibbs Sampler technique (Geman and Geman, 1984; Gelfand et al., 1990, Gelfand and Smith, 1990) to find the posterior distributions of the parameters and the random effects in the context of generalized linear models with random effects. The strength of Bayes approach lies in its flexibility in assessing the uncertainty in the random effects and functions of model parameters (Breslow and Clayton, 1993). The greatest advantage of Gibbs sampler is its ease of implementation. However, it is computationally intensive. Moreover, Hobert and Casella (1996) and Natarajan and McCulloch (1995) found that for models with random effects the posterior distribution of the parameters may not exist for diffuse priors, but that this problem may not be detected while computing, and thus wrong estimates can result (McCulloch, 1997).

Full Likelihood Approach

Anderson and Aitkin (1985), Hedeker and Gibbons (1994) and McCulloch (1997) approached the problem using a full likelihood approach. Anderson and Aitkin (1985) used Gaussian quadrature to approximate the integral in using maximum likelihood estimation in the logistic model with a single random effect. Hedeker and Gibbons (1994) also used Gauss-Hermite quadrature technique to find the marginal maximum likelihood estimators in ordinal regression models with multiple random effects. The advantage of Gaussian quadrature technique is that the precision of the estimation can be improved by increasing the number of quadrature points. However, as the number of

random effects increases, the number of quadrature points that have to be summed over increases exponentially, and so will the computational time. However, Bock, Gibbons and Muraki (1988) pointed out that the number of points for each random effect can be reduced as the number of random effects increases, without hurting the accuracy of the approximation.

McCulloch (1997) adapted the Monte Carlo version of the EM algorithm (MCEM) (Tanner 1993; Ledholter and Chan, 1994) for use in generalized linear models with random effects by incorporating a Metropolis-Hastings step. He also proposed a Monte Carlo version of the Newton-Raphson algorithm (MCNR) and improved the performance of simulated maximum likelihood developed by Geyer and Thompson (1992) and Gelfand and Charlin (1993) by preceding it with MCEM or MCNR. He compared these methods with the penalized quasi-likelihood approach using simulated data with large variance, which is known to be where the penalized quasi-likelihood suffers serious downward bias. The three methods were found to perform better than the penalized quasi-likelihood. However, the Monte Carlo methods have the same problem as the Gaussian quadrature technique in that the estimation takes time. Besides, the convergence is stochastic. That is, when the iterations converge, the convergent value will vary randomly within a small range of the maximum likelihood estimate. (Chan and Ledholter, 1994). This produces problems of deciding whether the MCEM or MCNR has really converged.

Quasi-Likelihood / Approximate Likelihood Approach

Goldstein (1991) and Longford (1984, 1988a) arrived at the same results via different routes (Goldstein, 1991; Rodriguez and Goldman, 1995). Goldstein (1991) completely avoided marginal likelihood estimation. He used the linearized dependent variable (McCullagh and Nelder, 1989) to borrow the strength of the normal theory methodology and proposed iterative generalized least squares to do the computation. Longford (1994, 1988) arrived at his approximation to the marginal likelihood integral by both using a second Taylor expansion around zero of the random effects and taking advantage of the normal theory. Breslow and Clayton (1993) considered such approaches as a marginal quasi-likelihood approach (MQL) because the conditional expectation in both cases is expanded around zero for the random effects. However, Rodriguez and Goldman (1995) conducted simulations on both packages as well as Goldstein's second order MQL (1991) and found the estimates to suffer substantial downward bias when the variances of the random effects are large.

Raudenbush (1993) extended Stiratelli, Laird, and Ware's (1984) posterior modal approach for binary responses to generalized linear models with random effects and also improved the efficiency of the approach by adopting Schall's framework (1991). He also used the linearized dependent variable with the conditional expectation expanded around the current estimates of both the random and fixed effects. As a result, although motivated in seemingly very different ways, the estimating equations used by Breslow and Clayton (1993) and Raudenbush (1993) are the same. Nevertheless, the fixed effects

and variance estimates also suffer underestimation (Yang, 1994), which, though not as severe as in the case of MQL, can still be serious when the variance of the random effects is large.

Breslow and Clayton (1993) used Laplace's method to derive the score function of the penalized quasi-likelihood (PQL) derived by Stiratelli et al. (1984). For the fixed and random effects estimation, they modified Green's (1987) Fisher scoring for estimating equations so as to borrow the strength of the normal theory linear model. For estimation of the variance components, they derived estimating equations, again using the normal theory, from the "REML version" of the profile likelihood of the approximate marginal likelihood, ignoring the dependence between the fixed effects and the variance. However, they showed PQL to be downward biased for estimates of both fixed effects and the variance components (Breslow and Clayton, 1993).

In an attempt to asymptotically correct the biases in approximate estimators of regression coefficients and the variance in generalized linear models with a single variance, Breslow and Lin (1995) expanded the joint distribution, using Taylor series, of the data and the random effects to the second and fourth orders around the current estimates, and then used Laplace's method to approximate the marginal likelihood. They termed the approximation up to second-order Taylor expansion the "first-order Laplace approximation" and that up to the fourth-order Taylor expansion the "second-order Laplace approximation". They found that the first order Laplace approximation for the variance estimator was seriously biased while the second order Laplace approximation

was better. Lin and Breslow (1996) extended the approximation further into models with multiple components of dispersion, with zero correlations among them.

Given the computational burden and other problems for Bayes and full likelihood approaches, the approximate/quasi- likelihood approach seems to be worth exploring more. The idea of Breslow and Lin (1995) and Lin and Breslow (1996) provides a way for the approximate/quasi -likelihood approach to better approximate the marginal likelihood and thus the estimation. Working in the approximate likelihood paradigm and trying to reduce the underestimation, this dissertation will extend their idea to the most general case where multilevel logistic models will have arbitrary number of random effects with a general variance covariance matrix.

Chapter 3

METHOD

Introduction

This chapter will find the approximate marginal log-likelihood using Laplace's method. Then, it will find the derivative of each term in the approximate log-likelihood in order to apply the approximate Fisher scoring for the estimation of the fixed effects and the variance-covariance components of the random effects.

Following is a list of all the formulae from Magnus and Neudecker (1988) and Magnus (1988) that are needed for the derivation. In proving theories in this section, these formulae will be referred to only by equation numbers.

$$vec(ABC) = (C^T \otimes A)vecB \qquad \text{(F1)}$$

$$vec(A)^T vec(B) = tr(A^T B) \qquad \text{(F2)}$$

$$(A \otimes B)(C \otimes D) = AC \otimes BD \qquad \text{(F3)}$$

$$tr(A \otimes B) = tr(A)tr(B) \qquad \text{(F4)}$$

$$tr(Axx^T) = x^T Ax \qquad \text{(F5)}$$

$$d\log|F| = trF^{-1}dF\,, \qquad \text{(F6)}$$

$$trAdX = (vecA^T)^T vecdX\,,\ A \text{ being constant} \qquad \text{(F7)}$$

$vecdAXB = vecA(dX)B = (B^T \otimes A)vecdX$, A and B being constants (F8)

$vecdX^{-1} = -((X^T)^{-1} \otimes X^{-1})vecdX$ (F9)

$[vec(ABC)]^T = [(C^T \otimes A)vecB]^T = (vecB)^T(C^T \otimes A)^T = (vecB)^T(C \otimes A^T)$ (F10)

$(b^T \otimes a^T) = \left(vec(ab^T)\right)^T$, b being a column vector. (F11)

The Logistic Model

We consider dichotomous responses from individuals nested within group *i*: $y_i = \mu_i + e_i$, where y_i is an $n_i \times 1$ response vector of either 1 or 0 for cluster i, with elements y_{ij}, i ranging from 1 to I, and j ranging from 1 to n_i. e_i is an $n_i \times 1$ column vector of error terms. μ_i is an $n_i \times 1$ column vector, the conditional mean of the *i*th cluster given b_i, each element being

$$\mu_{ij} = E\left(y_{ij} = 1 \middle| b_i\right) = \frac{1}{1 + \exp(-X_{ij}^T\beta - Z_{ij}^T b_i)} = \frac{1}{1 + \exp(-\eta_{ij})}. \quad (1)$$

Thus, each term μ_{ij} in μ_i is related to each term η_{ij} in η_i through the link function

$$\eta_{ij} = g(\mu_{ij}) = \log\left(\frac{\mu_{ij}}{1 - \mu_{ij}}\right). \quad (2)$$

$\eta_i = X_i\beta + Z_i b_i$ is a column vector, the linear predictor of the *i*th cluster. Here X_i is an $n_i \times p$ design matrix for the fixed effects, β, of the *i*th cluster, β being a $p \times 1$ vector. Z_i is an $n_i \times q$ design matrix for the random effects, b_i, of the *i*th cluster, b_i being a $q \times 1$ column vector that has a distribution $N(0, D)$, where D is a $q \times q$ variance-covariance matrix of b.

Submodel

If $D = \theta$, θ a scalar, then b_i is a scalar, and Z_i becomes an $n_i \times 1$ vector of 1. As a result, the conditional expectation becomes $\mu_{ij} = E(y_{ij} = 1|b_i) = \frac{1}{1+\exp(-X_{ij}^T\beta - b_i)}$, and $\eta_{ij} = X_{ij}^T\beta + b_i$.

Likelihood

For each observation j in the ith cluster, the conditional density of y_{ij} given b_i is

$$f(y_{ij}|b_i) = \mu_{ij}^{y_{ij}}(1-\mu_{ij})^{1-y_{ij}}. \quad (3)$$

Then for the ith cluster, the conditional density is

$$f(y_i|b_i) = \prod_{j=1}^{n_i} \mu_{ij}^{y_{ij}}(1-\mu_{ij})^{1-y_{ij}}. \quad (4)$$

The log of Equation 4 is $l_i = y_i^T\eta_i + s_i$, where $s_i = \sum_{j=1}^{n_i} s_{ij}$, with $s_{ij} = \log(1-\mu_{ij})$. Thus, the conditional log-likelihood of all the clusters is

$$\sum_{i=1}^{I} l_i = \sum_{i=1}^{I} y_i^T\eta_i + \sum_{i=1}^{I} s_i. \quad (5)$$

To get the marginal likelihood, we wish to integrate out b_i from the conditional density of y_{ij}:

$$L = \int\prod_i f(y_i|b_i)p(b)db_i = \prod_i \frac{1}{(2\pi)^{q/2}}|D|^{-1/2}\int\exp(l_i - \frac{1}{2}b_i^T D^{-1}b_i)db_i. \quad (6)$$

Likelihood of the Submodel

For $D = \theta$, the joint density of y_{ij} and b_i is

$$L = \int\prod_i f(y_i|b_i)p(b)db_i = \prod_i \frac{1}{(2\pi\theta)^{1/2}} \int \exp(l_i - \frac{1}{2\theta} b_i^2)db_i .$$

Approximation to the Log-Likelihood

Direct integration of Equation 6 is impossible and numerical integration cumbersome. Breslow and Lin (1995) used Laplace's method to approximate the integral, limiting the variance of the random effect(s) to either a scalar (Breslow and Lin, 1995) or a diagonal matrix (Lin and Breslow, 1996), however. That is, they first approximated the integrand in Equation 6 using Taylor expansion. Then by regarding the second-order term of the Taylor series as the kernel of a normal distribution, $N(\tilde{b}_i, -(\tilde{l}_i^{(2)} - D^{-1})^{-1})$, with $\tilde{l}_i^{(2)}$ being the second derivative of l_i evaluated at $\tilde{b}_i$, they took expectation of the other terms in the series under this new normal density, and approximated the integral as a sum of moments (up to the fourth moment). Note that $Q_i = -(\tilde{l}_i^{(2)} - D^{-1})^{-1}$ is also the posterior variance of b_i given $y_i, D,$ and β, if the linearized dependent variable (McCullagh and Nelder, 1989) y_i^* is assumed to be independently and identically normally distributed, i.e.,

$$y_i^* \sim N(X_i\beta + Z_i b_i,\ Z_i^T D Z_i + W_i).$$

In this section, we will generalize their approach to allow covariances among random effects. In addition, we will improve the accuracy of the approximation by including terms up to the sixth moment.

First, we approximate $l_i - \frac{1}{2} b_i^T D^{-1} b_i$ inside the exponent of the integrand with the Multivariate Taylor expansion around a current estimate of b_i, $\tilde{b}_i$, that maximizes the expansion. (See Appendix A: Multivariate Taylor Series Expansion)

$$l_i - \frac{1}{2} b_i^T D^{-1} b_i \approx \tilde{l}_i - \frac{1}{2} \tilde{b}_i^T D^{-1} \tilde{b}_i + \frac{\partial l_i}{\partial b^T}(b_i - \tilde{b}_i) - \tilde{b}_i^T D^{-1}(b_i - \tilde{b}_i)$$

$$+ \frac{1}{2}(b_i - \tilde{b}_i)^T \left(\frac{\partial^2 l_i}{\partial b \partial b^T} - D^{-1} \right)(b_i - \tilde{b}_i)$$

$$+ \frac{1}{3!}\left[(b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T\right] \frac{\partial \, vec\left(\frac{\partial^2 l_i}{\partial b \partial b^T} \right)}{\partial b^T}(b_i - \tilde{b}_i)$$

$$+ \frac{1}{4!}\left[(b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T\right] \frac{\partial \, vec\left(\frac{\partial \, vec\left(\frac{\partial^2 l_i}{\partial b \partial b^T} \right)}{\partial b^T} \right)}{\partial b^T}(b_i - \tilde{b}_i)$$

$$+ \frac{1}{5!}\left[(b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T\right] \times$$

$$\frac{\partial \, vec\left(\frac{\partial \, vec\left(\frac{\partial^2 l_i}{\partial b \partial b^T} \right)}{\partial b^T} \right)}{\partial \, vec \frac{\partial b^T}{\partial b^T}}(b_i - \tilde{b}_i)$$

$$+ \frac{1}{6!}\left[(b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T\right] \times$$

$$\frac{\partial\, vec\left(\dfrac{\partial\, vec\left(\dfrac{\partial^2 l_i}{\partial\, b\partial\, b^T}\right)}{\partial\, b^T}\right)}{\partial\, vec\dfrac{\partial\, b^T}{\partial\, b^T}}\cdot\frac{\partial\, vec\dfrac{\partial\, b^T}{\partial\, b^T}}{\partial\, vec\dfrac{\partial\, b^T}{\partial\, b^T}}(b_i-\tilde{b}_i)$$

where

1. $\tilde{l}_i$ is the value of $l_i = y_i^T\eta_i + s_i$, evaluated at the current estimates β, D and $\tilde{b}_i$;

2. $\dfrac{\partial\, l_i}{\partial\, b^T} = \tilde{l}_i^{(1)T} = (y_i-\mu_i)^T Z_i = (y_i^*-\eta_i)^T W_i Z_i$, with $\tilde{l}_i^{(1)} = Z_i^T W_i(y_i^*-\eta_i)$, where W_i is $diag[w_{ij}]$, with $w_{ij} = \mu_{ij}(1-\mu_{ij})$, the derivative of μ_{ij} with respect to η_{ij}, and $y_i^* = W_i^{-1}(y-\mu_i)+\eta_i$, the 'linearized dependent variable' (McCullagh & Nelder, 1989).

3. $\dfrac{\partial^2 l_i}{\partial b\partial b^T} = \dfrac{\partial \tilde{l}_i^{(1)}}{\partial b\partial b^T} = \tilde{l}_i^{(2)} = -Z_i^T W_i Z_i$.

4. $\dfrac{\partial vec\left(\dfrac{\partial^2 l_i}{\partial b\partial b^T}\right)}{\partial b^T} = \dfrac{\partial vec\tilde{l}_i^{(2)}}{\partial b^T} = \tilde{l}_i^{(3)} = -(Z_i^T\otimes Z_i^T)A_i Z_i$, where, $A_i = \sum_{j=1}^{n_i} a_{ij}(E_{ij}E_{ij}^T\otimes E_{ij})$, an $n_i^2\times n_i$ matrix, with $a_{ij} = \mu_{ij}(1-\mu_{ij})(1-2\mu_{ij}) = w_{ij}(1-2\mu_{ij})$, the second derivative of μ_{ij} with respect to η_{ij}, and E_{ij} an $n_i\times 1$ vector with the j th entry being 1, the others being 0.

5. $\dfrac{\partial vec\left\{\dfrac{\partial vec\left(\dfrac{\partial^2 l_i}{\partial b\partial b^T}\right)}{\partial b^T}\right\}}{\partial b^T} = \dfrac{\partial vec\tilde{l}_i^{(3)}}{\partial b^T} = \tilde{l}_i^{(4)} = -Z_i^T\otimes(Z_i^T\otimes Z_i^T)G_i Z_i$

where $G_i = \sum_{j=1}^{n_i} g_{ij}(E_{ij}E_{ij}^T \otimes E_{ij} \otimes E_{ij})$, an $n_i^3 \times n_i$ matrix, with $g_{ij} = w_{ij}(1-6w_{ij})$, the third derivative of μ_{ij} with respect to η_{ij}.

$$6.\quad \frac{\partial vec\left\{\dfrac{\partial vec\left(\dfrac{\partial^2 l_i}{\partial b \partial b^T}\right)}{\partial b^T}\right\}^T}{\partial b^T} = \tilde{l}_i^{(5)} = -(Z_i^T \otimes Z_i^T \otimes Z_i^T \otimes Z_i^T)H_i Z_i$$

where $H_i = \sum_{j=1}^{n_i} h_{ij}(E_{ij}E_{ij}^T \otimes E_{ij} \otimes E_{ij} \otimes E_{ij})$, an $n_i^4 \times n_i$ matrix, with $h_{ij} = a_{ij}(1-12w_{ij})$, the third derivative of μ_{ij} with respect to η_{ij}.

$$7.\quad \frac{\partial vec\left\{\dfrac{\partial vec\left(\dfrac{\partial^2 l_i}{\partial b \partial b^T}\right)}{\partial b^T}\right\}^T}{\partial b^T} = \tilde{l}_i^{(6)} = -(Z_i^T \otimes Z_i^T \otimes Z_i^T \otimes Z_i^T \otimes Z_i^T)F_i Z_i$$

where $F_i = \sum_{j=1}^{n_i} f_{ij}(E_{ij}E_{ij}^T \otimes E_{ij} \otimes E_{ij} \otimes E_{ij} \otimes E_{ij})$, an $n_i^5 \times n_i$ matrix, with

$f_{ij} = g_{ij}(1-12w_{ij}) - 12a_{ij}^2$, the fifth derivative of μ_{ij} with respect to η_{ij}.

All of the above are derived by using Equation F8, regarding the matrix to be differentiated as X and those on the two sides as A and B.

Then we will prove that the approximate log-likelihood is approximately

$$\log(L) \approx L_6 = \sum_{i=1}^{I}\left\{\frac{-1}{2}\log|D| - \frac{1}{2}\log|Q_i^{-1}| + \tilde{l}_i - \frac{1}{2}\tilde{b}_i^T D^{-1}\tilde{b}_i + \log M_i\right\}, \qquad (7)$$

where $Q_i^{-1} = -(\tilde{l}_i^{(2)} - D^{-1}) = \left(D^{-1} + Z_i^T W_i Z_i\right)$, $B_{ij} = Z_{ij}^T Q_i Z_{ij}$, and

$$M_i = 1 - \frac{1}{8}\sum_j^{n_i} g_{ij} B_{ij}^2 - \frac{1}{48}\sum_j^{n_i} r_{ij} B_{ij}^3 + \frac{15}{72} k_i^T Q_i k_i, \text{ with } k_i = \left(\sum_k^{n_i} a_{ik} B_{ik} Z_{ik}\right).$$

Proof Substituting the integrand with its Taylor expansion, the marginal likelihood becomes

$$\begin{aligned} L \approx \prod_{i=1}^{I}(2\pi)^{\frac{-1}{2}}|D|^{\frac{-1}{2}} \int \exp\Big\{ & \tilde{l}_i - \frac{1}{2}\tilde{b}_i^T D^{-1}\tilde{b}_i + (\tilde{l}_i^{(1)} - D^{-1}\tilde{b}_i)^T (b_i - \tilde{b}_i) \\ & + \frac{1}{2}(b_i - \tilde{b}_i)^T (\tilde{l}_i^{(2)} - D^{-1})(b_i - \tilde{b}_i) + \frac{1}{3!}\left[(b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T\right]\tilde{l}_i^{(3)}(b_i - \tilde{b}_i) \\ & + \frac{1}{4!}\left[(b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T\right]\tilde{l}_i^{(4)}(b_i - \tilde{b}_i) \\ & + \frac{1}{5!}\left[(b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T\right]\tilde{l}_i^{(5)}(b_i - \tilde{b}_i) \qquad (8) \\ & + \frac{1}{6!}\left[(b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T\right]\tilde{l}_i^{(6)}(b_i - \tilde{b}_i)\Big\} db_i, \end{aligned}$$

where $\tilde{b}_i$ is the maximizing value for L, i.e., $\tilde{b}_i$ solves $\tilde{l}_i^{(1)} - D^{-1}\tilde{b}_i = 0$. That is, $\tilde{b}_i = D\tilde{l}_i^{(1)} = DZ_i^T W_i(y_i^* - X_i\beta - Z_i\tilde{b}_i)$. To find $\tilde{b}_i$, collect $\tilde{b}_i$ at the left hand side, $(I + DZ_i^T W_i Z_i)\tilde{b}_i = DZ_i^T W_i(y_i^* - X_i\beta)$. Thus, $\tilde{b}_i = (I + DZ_i^T W_i Z_i)^{-1} DZ_i^T W_i(y_i^* - X_i\beta) = Q_i Z_i^T W_i(y_i^* - X_i\beta)$. Consequently, $\tilde{l}_i^{(1)} - D^{-1}\tilde{b}_i$ vanishes. The second order term is retained as the kernel of the new Normal density $N(\tilde{b}_i,\ Q_i)$. Next we consider the approximation of the higher order terms.

For $R = \frac{1}{3!}\left[(b_i - \tilde{b}_i)^T (b_i - \tilde{b}_i)^T\right]\tilde{l}_i^{(3)}(b_i - \tilde{b}_i)$

$$+\frac{1}{4!}\left[(b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T\right]\tilde{l}_i^{(4)}(b_i - \tilde{b}_i)$$

$$+\frac{1}{5!}\left[(b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T\right]\tilde{l}_i^{(5)}(b_i - \tilde{b}_i)$$

$$+\frac{1}{6!}\left[(b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T\right]\tilde{l}_i^{(6)}(b_i - \tilde{b}_i)$$

in the exponent, we adopt the formula for approximating an exponential:

$\exp(R) \approx 1 + R + \frac{1}{2}R^2 + \frac{1}{3!}R^3 + \ldots$. However, through experiments, the contribution of the expansion terms to the log-likelihood of orders higher than the seventh is negligible. We only take up to R and $\frac{R^2}{2}$. Moreover, we find the approximation $\frac{R^2}{2}$ for $\frac{1}{2}\left[\frac{1}{3!}\left((b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T\right)\tilde{l}_i^{(3)}(b_i - \tilde{b}_i)\right]^2$ to be non-negligible in approximating the log-likelihood. Therefore,

$$\exp\{R\} = \exp\left\{\frac{1}{3!}\left[vec(b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T\right]\tilde{l}_i^{(3)}(b_i - \tilde{b}_i)\right.$$

$$+\frac{1}{4!}\left[(b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T\right]\tilde{l}_i^{(4)}(b_i - \tilde{b}_i)$$

$$+\frac{1}{5!}\left[(b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T\right]\tilde{l}_i^{(5)}(b_i - \tilde{b}_i)$$

$$\left.+\frac{1}{6!}\left[(b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T\right]\tilde{l}_i^{(6)}(b_i - \tilde{b}_i)\right\}$$

$$\approx \left\{1 + \frac{1}{3!}\left[(b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T\right]^T \tilde{l}_i^{(3)}(b_i - \tilde{b}_i)\right.$$

$$+\frac{1}{4!}\left[(b_i-\tilde{b}_i)^T\otimes(b_i-\tilde{b}_i)^T\otimes(b_i-\tilde{b}_i)^T\right]\tilde{l}_i^{(4)}(b_i-\tilde{b}_i)$$

$$+\frac{1}{5!}\left[(b_i-\tilde{b}_i)^T\otimes(b_i-\tilde{b}_i)^T\otimes(b_i-\tilde{b}_i)^T\otimes(b_i-\tilde{b}_i)^T\right]\tilde{l}_i^{(5)}(b_i-\tilde{b}_i)$$

$$+\frac{1}{6!}\left[(b_i-\tilde{b}_i)^T\otimes(b_i-\tilde{b}_i)^T\otimes(b_i-\tilde{b}_i)^T\otimes(b_i-\tilde{b}_i)^T\otimes(b_i-\tilde{b}_i)^T\right]\tilde{l}_i^{(6)}(b_i-\tilde{b}_i)$$

$$+\frac{1}{2}\left(\frac{1}{3!}\left[(b_i-\tilde{b}_i)^T\otimes(b_i-\tilde{b}_i)^T\right]\tilde{l}_i^{(3)}(b_i-\tilde{b}_i)\right)^2\Bigg\}.$$ Then Equation 8 becomes

$$L\approx\prod_{i=1}^{m}(2\pi)^{\frac{-1}{2}}|D|^{\frac{-1}{2}}\exp(\tilde{l}_i-\frac{1}{2}\tilde{b}_i^TD^{-1}\tilde{b}_i)\int\Bigg\{\exp\left[\frac{1}{2}(b_i-\tilde{b}_i)^T(\tilde{l}_i^{(2)}-D^{-1})(b_i-\tilde{b}_i)\right]\times$$

$$\Bigg\langle 1+\frac{1}{3!}\left[(b_i-\tilde{b}_i)^T\otimes(b_i-\tilde{b}_i)^T\right]\tilde{l}_i^{(3)}(b_i-\tilde{b}_i)$$

$$+\frac{1}{4!}\left[(b_i-\tilde{b}_i)^T\otimes(b_i-\tilde{b}_i)^T\otimes(b_i-\tilde{b}_i)^T\right]\tilde{l}_i^{(4)}(b_i-\tilde{b}_i)$$

$$+\frac{1}{5!}\left[(b_i-\tilde{b}_i)^T\otimes(b_i-\tilde{b}_i)^T\otimes(b_i-\tilde{b}_i)^T\otimes(b_i-\tilde{b}_i)^T\right]\tilde{l}_i^{(5)}(b_i-\tilde{b}_i)$$

$$+\frac{1}{6!}\left[(b_i-\tilde{b}_i)^T\otimes(b_i-\tilde{b}_i)^T\otimes(b_i-\tilde{b}_i)^T\otimes(b_i-\tilde{b}_i)^T\otimes(b_i-\tilde{b}_i)^T\right]\tilde{l}_i^{(6)}(b_i-\tilde{b}_i)$$

$$+\frac{1}{2}\left(\frac{1}{3!}\left[(b_i-\tilde{b}_i)^T\otimes(b_i-\tilde{b}_i)^T\right]\tilde{l}_i^{(3)}(b_i-\tilde{b}_i)\right)^2\Bigg\rangle\Bigg\}db_i. \qquad (9)$$

Using Laplace's approximation method, multiplied by the normalizing constant, $(2\pi)^{\frac{1}{2}}|Q_i|^{\frac{1}{2}}$, Equation 9 becomes the expectation of 1 plus the third, fourth, fifth and sixth moments of a multivariate normal distribution with mean 0 and variance Q_i multiplied by $\tilde{l}_i^{(3)}$, $\tilde{l}_i^{(4)}$, $\tilde{l}_i^{(5)}$, $\tilde{l}_i^{(6)}$ and $(\tilde{l}_i^{(3)})^2$ respectively. However, the odd moment of a normal distribution is 0. Thus the likelihood can be approximated as (See APPENDIX B: The

Six Moments of a Multivariate Normal Distribution, APPENDIX C: Proof of the Substitution, and APPENDIX D: The Expectation of the Third Term Squared)

$$L \approx \prod_{i=1}^{m} (2\pi)^{\frac{-1}{2}} |D|^{\frac{-1}{2}} (2\pi)^{\frac{1}{2}} |Q_i|^{\frac{1}{2}} \exp(\tilde{l}_i - \frac{1}{2}\tilde{b}_i^T D^{-1}\tilde{b}_i) \times$$

$$E\Big\langle 1 + \frac{1}{24}\big[(b_i - \tilde{b}_i)^T \otimes (b - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T\big]\tilde{l}_i^{(4)}(b_i - \tilde{b}_i)$$

$$+ \frac{1}{720}\big[(b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T\big]\tilde{l}_i^{(6)}(b_i - \tilde{b}_i)$$

$$+ \frac{1}{72}\Big(\big[(b_i - \tilde{b}_i)^T \otimes (b_i - \tilde{b}_i)^T\big]\tilde{l}_i^{(3)}(b_i - \tilde{b}_i)\Big)^2\Big\rangle$$

$$= \prod_{i=1}^{m} |D|^{\frac{-1}{2}} |Q_i|^{\frac{1}{2}} \exp(\tilde{l}_i - \frac{1}{2}\tilde{b}_i^T D^{-1}\tilde{b}_i)\Big\langle 1 + \frac{1}{8}\{vec[Q_i \otimes Q_i]\}^T vec\tilde{l}_i^{(4)}$$

$$+ \frac{1}{48}\{vec[Q_i \otimes Q_i \otimes Q_i]\}^T vec\tilde{l}_i^{(6)} + \frac{15}{72}k_i^T Q_i k_i \Big\rangle.$$ The matrix pre-multiplying $vec\tilde{l}_i^{(4)}$ is the result of $\frac{1}{4!}$ multiplied by the fourth moment of the normal distribution. The matrix pre-multiplying $vec\tilde{l}_i^{(6)}$ is the result of $\frac{1}{6!}$ multiplied by the sixth moment of the normal distribution. The derivation of the last term can be found in APPENDIX D. Therefore, the approximate log-likelihood is

$$\log(L) \approx \sum_{i=1}^{I}\Big\{-\frac{1}{2}\log|D| - \frac{1}{2}\log|Q_i^{-1}| + \tilde{l}_i - \frac{1}{2}\tilde{b}_i^T D^{-1}\tilde{b}_i + \log\Big\langle 1 + \frac{1}{8}[vec[Q_i \otimes Q_i]]^T vec\tilde{l}_i^{(4)}$$

$$+ \frac{1}{48}\{vec[Q_i \otimes Q_i \otimes Q_i]\}^T vec\tilde{l}_i^{(6)} + \frac{15}{72}k_i^T Q_i k_i \Big\rangle\Big\}. \qquad (10)$$

Substituting $\tilde{l}_i^{(2)}$, $\tilde{l}_i^{(4)}$ $\tilde{l}_i^{(6)}$ and $\tilde{l}_i^{(3)}$ into Equation 10 gives

$$\log(L) \approx \sum_{i=1}^{I}\left\{\frac{-1}{2}\log|D| - \frac{1}{2}\log\left|Q_i^{-1}\right| + \tilde{l}_i - \frac{1}{2}\tilde{b}_i^T D^{-1}\tilde{b}_i + \right.$$

$$\log\left\langle 1 - \frac{1}{8}[vec(Q_i \otimes Q_i)]^T vec[(Z_i^T \otimes Z_i^T \otimes Z_i^T)G_i Z_i]\right.$$

$$-\frac{1}{48}\{vec[Q_i \otimes Q_i \otimes Q_i]\}^T vec[(Z_i^T \otimes Z_i^T \otimes Z_i^T \otimes Z_i^T \otimes Z_i^T)F_i Z_i]$$

$$\left.\left. + \frac{15}{72}k_i^T Q_i k_i \right\rangle\right\}. \qquad (11)$$

Now we simplify the fourth order term in Equation 11. Let

$e_i = [vec(Q_i \otimes Q_i]^T vec[(Z_i^T \otimes Z_i^T \otimes Z_i^T)G_i Z_i]$, ignoring the constant, $\frac{-1}{8}$. First, by Equation F1, regarding G_i as B in the formula,

$$vec[(Z_i^T \otimes Z_i^T \otimes Z_i^T)G_i Z_i] = vec[(Z_i^T \otimes Z_i^T \otimes Z_i^T)\sum_{j=1}^{J} g_{ij}(E_{ij}E_{ij}^T \otimes E_{ij} \otimes E_{ij})Z_i]$$

$$= vec[\sum_{j=1}^{J} g_{ij}(Z_i^T \otimes Z_i^T \otimes Z_i^T)(E_{ij}E_{ij}^T \otimes E_{ij} \otimes E_{ij})(Z_i \otimes 1 \otimes 1)]$$

$$= \sum_{j=1}^{J} g_{ij} vec[(Z_{ij}Z_{ij}^T \otimes Z_{ij} \otimes Z_{ij})]. \qquad (12)$$

Here 1 is a scalar 1. Z_{ij} is a $q \times 1$ column vector of the random effect design matrix for person j in group i. Since Equation 12 is the vectorization of Kronecker products of the same vector Z_{ij}, it can be re-written as

$$vec[(Z_i^T \otimes Z_i^T \otimes Z_i^T)G_i Z_i] = \sum_{j=1}^{J} g_{ij} vec[(Z_{ij}Z_{ij}^T \otimes Z_{ij}Z_{ij}^T)] \qquad (13)$$

(by F11, see also PA-12 in PRE-APPENDIX). As a result,

$$e_i = [vec[Q_i \otimes Q_i]]^T \sum_j^{n_i} g_{ij} vec(Z_{ij} Z_{ij}^T \otimes Z_{ij} Z_{ij}^T). \quad (14)$$

Since Q_i is symmetric, so is $Q_i \otimes Q_i$. By applying Equation F2, Equation 14 becomes

$e_i = \sum_j^{n_i} g_{ij} tr\{[Q_i \otimes Q_i](Z_{ij} Z_{ij}^T \otimes Z_{ij} Z_{ij}^T)\}$. Furthermore, by Equations F3 and F4, e_i is

simplified even more to $e_i = \sum_j^{n_i} g_{ij} \{tr[(Q_i Z_{ij} Z_{ij}^T]\}^2$. (15)

To get rid of the trace function of the above, we take advantage of Z_{ij} being a

$q \times 1$ vector, and use Equation F5. Therefore, Equation 14 becomes, finally,

$$e_i = \sum_j^{n_i} g_{ij} [Z_{ij}^T Q_i Z_{ij}]^2 = \sum_j^{n_i} g_{ij} B_{ij}^2 . \quad (16)$$

The sixth order term can be simplified in exactly the same way. That is ,

$$vec[(Z_i^T \otimes Z_i^T \otimes Z_i^T \otimes Z_i^T \otimes Z_i^T) F_i Z_i]$$

$$= vec[(Z_i^T \otimes Z_i^T \otimes Z_i^T \otimes Z_i^T \otimes Z_i^T) \sum_{j=1}^{J} f_{ij} (E_{ij} E_{ij}^T \otimes E_{ij} \otimes E_{ij} \otimes E_{ij} \otimes E_{ij}) Z_i]$$

$$= vec[\sum_{j=1}^{J} f_{ij} (Z_{ij} Z_{ij}^T \otimes Z_{ij} \otimes Z_{ij} \otimes Z_{ij} \otimes Z_{ij})]$$

$$= vec[\sum_{j=1}^{J} f_{ij} (Z_{ij} Z_{ij}^T \otimes Z_{ij} Z_{ij}^T \otimes Z_{ij} Z_{ij}^T)] \quad (17)$$

As a result, $q_i = \{vec[Q_i \otimes Q_i \otimes Q_i]\}^T \; vec[(Z_i^T \otimes Z_i^T \otimes Z_i^T \otimes Z_i^T \otimes Z_i^T) F_i Z_i]$

$$= \{vec[Q_i \otimes Q_i \otimes Q_i]\}^T \sum_j^{n_i} f_{ij} vec(Z_{ij} Z_{ij}^T \otimes Z_{ij} Z_{ij}^T \otimes Z_{ij} Z_{ij}^T)$$

$$= \sum_j^{n_i} f_{ij} tr[Q_i Z_{ij} Z_{ij}^T \otimes Q_i Z_{ij} Z_{ij}^T \otimes Q_i Z_{ij} Z_{ij}^T]$$

$$= \sum_j^{n_i} f_{ij} \left(tr[Q_i Z_{ij} Z_{ij}^T] \right)^3 = \sum_j^{n_i} f_{ij} \left(Z_{ij}^T Q_i Z_{ij} \right)^3 = \sum_j^{n_i} f_{ij} B_{ij}^3 \qquad (17)$$

Putting $\frac{-1}{8} e_i$ and $\frac{-1}{48} q_i$ back to the approximate log-likelihood, it becomes

$$\log(L) \approx L_6 = \sum_{i=1}^{I} \left\{ \frac{-1}{2} \log|D| - \frac{1}{2} \log|Q_i^{-1}| + \tilde{l}_i - \frac{1}{2} \tilde{b}_i^T D^{-1} \tilde{b}_i \right.$$

$$\left. + \log\left\langle 1 - \frac{1}{8} \sum_j^{n_i} g_{ij} B_{ij}^2 - \frac{1}{48} \sum_j^{n_i} f_{ij} B_{ij}^3 + \frac{15}{72} k_i^T Q_i k_i \right\rangle \right\}.$$

Hence we have finished the proof.

Approximation to the Submodel

For $D = \theta$, the log-likelihood of the data is obtained by substituting D by θ, Z_{ij} by the scalar 1. Therefore, $\tilde{l}_i^{(2)} = \sum_j^{n_i} \tilde{l}_{ij}^{(2)}$, $\tilde{l}_i^{(3)} = \sum_j^{n_i} \tilde{l}_{ij}^{(3)}$, $\tilde{l}_i^{(4)} = \sum_j^{n_i} \tilde{l}_{ij}^{(4)}$, and $\tilde{l}_i^{(6)} = \sum_j^{n_i} \tilde{l}_{ij}^{(6)}$ for the univariate case. Then we have

$$\log(L) \approx L_6 = \sum_{i=1}^{I} \left\{ \frac{-1}{2} \log\theta - \frac{1}{2} \log(\theta^{-1} - \sum_j^{n_i} \tilde{l}_{ij}^{(2)}) + \tilde{l}_i - \frac{1}{2\theta} \tilde{b}_i^2 \right.$$

$$+ \log\left\langle 1 + \frac{1}{8} \sum_j^{n_i} \tilde{l}_{ij}^{(4)} [(\theta^{-1} - \sum_j^{n_i} \tilde{l}_{ij}^{(2)})^{-1}]^2 + \frac{1}{48} \sum_j^{n_i} \tilde{l}_{ij}^{(6)} [(\theta^{-1} - \sum_j^{n_i} \tilde{l}_{ij}^{(2)})^{-1}]^3 \right.$$

$$\left. \left. + \frac{15}{72} \left[\sum_j^{n_i} \tilde{l}_{ij}^{(3)} \right]^2 [(\theta^{-1} - \sum_j^{n_i} l_{ij}^{(2)})^{-1}]^3 \right\rangle \right\}.$$

Approximate Fisher Scoring

This section will find the first partial derivatives of the terms in the approximate log-likelihood, L_6, with respect to the fixed effects β and the variance components D, separately, and then merge the derivatives into one score function vector, S_i, for each group *i*, so as to form the approximate Fisher scoring (Green, 1984),

$(\sum_i S_i S_i^T)^{-1} \sum_i S_i$ where $S_i = \begin{bmatrix} S_{\beta_i} \\ S_{D_i} \end{bmatrix}$. Note that to take the derivative of a scalar *s* with respect to a column vector *v* is the same as taking its derivative with respect to v^T, and then take the transpose of the resultant row vector. That is, $\frac{\partial s}{\partial v} = \left(\frac{\partial s}{\partial v^T}\right)^T$. The latter is used here because it is more straightforward in applying formulae by Magnus and Neudecker (1988).

Implicit Differential

The posterior mode of the random effects, $\tilde{b}_i$, depends on the variance D and the fixed effects β, i.e., $\tilde{b}_i = \tilde{b}_i(\beta, D) = D\tilde{l}_i^{(1)}$. In finding score functions for β and D we need to take this relation into account by finding the differentials, $\frac{\partial \tilde{b}_i}{\partial \beta^T}$ and $\frac{\partial \tilde{b}_i}{\partial (vecD)^T}$, through implicit differentiation.

$$\frac{\partial \tilde{b}_i}{\partial \beta^T} = \frac{\partial D\tilde{l}_i^{(1)}}{\partial \beta^T} = DZ_i^T \frac{\partial}{\partial \beta^T}(y_i - \tilde{\mu}_i) = -DZ_i^T W_i (X_i + Z_i \frac{\partial \tilde{b}_i}{\partial \beta^T}).$$

Collecting $\frac{\partial \tilde{b}_i}{\partial \beta^T}$ at the left hand side, $(I + DZ_i^T W Z_i)\frac{\partial \tilde{b}_i}{\partial \beta^T} = -DZ_i^T W_i X_i$.

Thus, pre-multiplying each side with D^{-1},

$$\frac{\partial \tilde{b}_i}{\partial \beta^T} = -(D^{-1} + Z_i^T W_i Z_i)^{-1} Z_i^T W_i X_i = -Q_i Z_i^T W_i X_i. \tag{19}$$

Then $\frac{\partial \eta_i}{\partial \beta^T} = \frac{\partial (X_i \beta + Z_i \tilde{b}_i)}{\partial \beta^T} = (X_i - Z_i Q_i Z_i^T W_i X_i)$. (20)

Similarly, $\frac{\partial \tilde{b}_i}{\partial (vecD)^T} = \frac{\partial D\tilde{l}_i^{(1)}}{\partial (vecD)^T} = [(y_i^* - \eta_i)^T W_i Z_i \otimes I_q] - DZ_i^T W_i (Z_i \frac{\partial \tilde{b}_i}{\partial (vecD)^T})$

$$\frac{\partial \tilde{b}_i}{\partial (vecD)^T} = (D^{-1} + Z_i^T W_i Z_i)^{-1} D^{-1} [(y_i^* - \eta_i)^T W_i Z_i \otimes I_q]. \tag{21}$$

$$\frac{\partial \eta_i}{\partial (vecD)^T} = \frac{\partial (X_i \beta + Z_i \tilde{b}_i)}{\partial (vecD)^T} = Z_i Q_i D^{-1} [(y_i^* - \eta_i)^T W_i Z_i \otimes I_q] \tag{22}$$
$$= [(y_i^* - \eta_i)^T W_i Z_i \otimes Z_i Q_i D^{-1}].$$

In the following derivatives for each term in L_6, all the terms that are functions of $\tilde{b}_i$, e.g., η_i, μ_i, W_i, A_i, G_i, and F_i, will have partial derivatives not only with respect to the apparent β and $vecD$ but also those inside $\tilde{b}_i$ using the above derivatives.

Score Function of the Fixed Effects

I will prove that the score function of β for group i is

$$S_\beta = \frac{-1}{2} \sum_j^{n_i} a_{ij} Z_{ij}^T Q_i Z_{ij} [X_{ij} - X_i^T W_i Z_i Q_i Z_{ij}] + X_i^T W_i (y_i^* - \eta_i)$$

$$+ \frac{1}{M_i} \Big(-\frac{1}{8} [\sum_k^{n_i} h_{ik} B_{ij}^2 - 2 \sum_k^{n_i} \sum_j^{n_i} g_{ij} B_{ij} a_{ik} (Z_{ij}^T Q_i Z_{ik})^2] [X_{ik} - X_i^T W_i Z_i Q_i Z_{ik}]$$

$$- \frac{1}{48} [\sum_k^{n_i} p_{ik} B_{ik}^3 - 3 \sum_k^{n_i} \sum_j^{n_i} f_{ij} B_{ij}^2 a_{ik} (Z_{ij}^T Q_i Z_{ik})^2] [X_{ik} - X_i^T W_i Z_i Q_i Z_{ik}]$$

$$+\frac{15}{36}\left(\sum_{j}^{n_i} g_{ij} k_i^T Q_i Z_{ij} Z_{ij}^T Q_i Z_{ij}\right)\left(X_{ij} - X_i^T W_i Z_i Q_i Z_{ij}\right)$$

$$-\frac{15}{36}\left[\sum_{j}^{n_i}\sum_{k}^{n_i} a_{ij} a_{ik} k_i^T Q_i Z_{ij} (Z_{ij}^T Q_i Z_{ik})^2\right]\left[X_{ik} - X_i^T W_i Z_i Q_i Z_{ik}\right]$$

$$-\frac{15}{72}\left[\sum_{k}^{n_i} a_{ik} (k_j^T Q_i Z_{ik})^2\right]\left[X_{ik} - X_i^T W_i Z_i Q_i Z_{ik}\right]\Bigg\},$$

where $p_{ik} = h_{ik}(1-12w_{ik}) - 36a_{ik}g_{ik}$.

Proof Making use of Equations F6 first, and then F7, and F8, the derivative of the second term in L_6 is:

$$\frac{\partial}{\partial \beta^T}\left(\frac{-1}{2}\log\left|D^{-1} + Z_i^T W_i Z_i\right|\right) = \frac{-1}{2} tr\left((D^{-1} + Z_i^T W_i Z_i)^{-1} \frac{\partial (Z_i^T W_i Z_i)}{\partial \beta^T}\right)$$

$$= \frac{-1}{2}[vec(D^{-1} + Z_i^T W_i Z_i)^{-1}]^T \frac{\partial\, vec(Z_i^T W_i Z_i)}{\partial \beta^T}$$

$$= \frac{-1}{2}\sum_{j}^{n_i} a_{ij}[vec Q_i]^T (Z_{ij} \otimes Z_{ij})[X_{ij}^T + Z_{ij}^T \frac{d\tilde{b}_i}{d\beta^T}]$$

$$= \frac{-1}{2}\sum_{j}^{n_i} a_{ij} Z_{ij}^T Q_i Z_{ij}[X_{ij}^T - Z_{ij}^T Q_i Z_i^T W_i X_i] \text{ (by Equations F2 and 19),} \qquad (23)$$

where $\frac{\partial vec(Z_i^T W_i Z_i)}{\partial \beta^T} = (Z_i^T \otimes Z_i^T) A_i (X_i - Z_i Q_i Z_i^T W_i X_i)$

$$= \sum_{j}^{n_i} a_{ij} (Z_{ij} \otimes Z_{ij}) (X_{ij}^T - Z_{ij}^T Q_i Z_i^T W_i X_i). \qquad (24)$$

Equation 24 is obtained by regarding Z_i^T as A in Equation F8, and Z_i as B. We vectorize W_i and take its derivative to get $A_i(X_i - Z_i Q_i Z_i^T W_i X_i)$. However,

$A_i = \sum_j^{n_i} a_{ij}(E_{ij}E_{ij}^T \otimes E_{ij})$. We make use of this special structure to decompose A_i and get Equation 23. The transpose of Equation 23 is the first term in the score function.

$$\frac{\partial}{\partial \beta^T}\tilde{l}_i = (y_i - \mu_i)^T(X_i - Z_iQ_iZ_i^TW_iX_i) = (y_i^* - \eta_i)^TW_i(X_i - Z_iQ_iZ_i^TW_iX_i) \quad (25)$$

The derivation is similar to that of the $\frac{\partial l_i}{\partial b^T}$ in the derivation of the approximate log-likelihood.

$$\frac{\partial}{\partial \beta^T}(-\frac{1}{2}\tilde{b}_i^TD^{-1}\tilde{b}_i) = -\tilde{b}_i^TD^{-1}\left(\frac{\partial \tilde{b}_i}{\partial \beta^T}\right) = \tilde{b}_i^TD^{-1}Q_iZ_i^TW_iX_i. \quad (26)$$

However, since $\tilde{b}_i = DZ_i^TW_i(y_i^* - \eta_i)$, $D^{-1}\tilde{b}_i - Z_i^TW_i(y_i^* - \eta_i) = 0$. In adding up Equations 25 and 26, $\tilde{b}_i^TD^{-1}Q_iZ_i^TW_iX_i - (y_i^* - \eta_i)W_iZ_iQ_iZ_i^TW_iX_i = 0$, with only $(y_i^* - \eta_i)^TW_iX_i$ left, the transpose of which is the second term in the score function.

The derivative of $\log M_i$ in the approximate log-likelihood function will be

$$\frac{\partial \log M_i}{\alpha \beta^T} = \frac{1}{M_i}\partial\frac{\partial M_i}{\partial \beta^T}. \quad (27)$$

We take derivative of the first term in M_i.

$$\frac{\partial}{\partial \beta^T}\sum_j^{n_i} g_{ij}B_{ij}^2 = \sum_j^{n_i} h_{ij}B_{ij}^2[X_{ij}^T - Z_{ij}^TQ_iZ_i^TW_iX_i]$$
$$-2\sum_j^{n_i} g_{ij}B_{ij}(Z_{ij}^TQ_i \otimes Z_{ij}^TQ_i)(Z_i^T \otimes Z_i^T)A_i(X_i - Z_iQ_iZ_i^TW_iX_i)$$
$$= [\sum_k^{n_i} h_{ik}B_{ik}^2 - 2\sum_k^{n_i}\sum_j^{n_i} g_{ij}B_{ij}a_{ik}(Z_{ij}^TQ_iZ_{ik})^2][X_{ik}^T - Z_{ik}^TQ_iZ_i^TW_iX_i] \quad (28)$$

where $B_{ij} = Z_{ij}^T Q_i Z_{ij}$, $B_{ij}^2 = (Z_{ij}^T Q_i Z_{ij})^2$, and h_{ij} is the derivative of g_{ij} with respect to η_{ij}. The first item in the right hand side of (28) results from taking derivative of g_{ij}. The second term comes from taking derivative of the B_{ij}^2. Since B_{ij}^2 is a scalar, differentiating the second term leads to $2\sum_j^{n_i} g_{ij} B_{ij} \frac{\partial B_{ij}}{\partial \beta^T}$, for which we use Equations F8 and F9. Applying Equation F9, we have $\frac{\partial}{\partial \beta^T} vec Q_i = -(Q_i \otimes Q_i) \frac{\partial vec(Z_i^T W_i Z_i)}{\partial \beta^T}$. (29)

Substituting Equation 24 inside Equation 29 and then applying F3 we have the second term in Equation 28. Then, the transpose of Equation 28 is

$$\frac{\partial}{\partial \beta} \sum_j^{n_i} g_{ij} B_{ij}^2 = [\sum_k^{n_i} h_{ik} B_{ik}^2 - 2\sum_k^{n_i}\sum_j^{n_i} g_{ij} B_{ij} a_{ik} (Z_{ij}^T Q_i Z_{ik})^2][X_{ik} - X_i^T W_i Z_i Q_i Z_{ik}]. \quad (30)$$

Multiplied by $\frac{-1}{8}$, Equation 30 is the contribution of the fourth order term of the Taylor series to the score function of β_i.

Similarly,

$$\frac{\partial}{\partial \beta^T} \sum_j^{n_i} f_{ij} B_{ij}^3 = [\sum_k^{n_i} p_{ik} B_{ik}^3 - 3\sum_k^{n_i}\sum_j^{n_i} f_{ij} B_{ij}^2 a_{ik} (Z_{ij}^T Q_i Z_{ik})^2][X_{ik}^T - Z_{ik}^T Q_i Z_i^T W_i X_i], \quad (31)$$

where $p_{ij} = h_{ij}(1 - 12 w_{ij}) - 36 a_{ij} g_{ij}$, the derivative of f_{ij} with respect to η_{ij}.

The transpose of Equation 31 is

$$\frac{\partial}{\partial \beta} \sum_j^{n_i} f_{ij} B_{ij}^3 = [\sum_k^{n_i} p_{ik} B_{ik}^3 - 3\sum_k^{n_i}\sum_j^{n_i} f_{ij} B_{ij}^2 a_{ik} (Z_{ij}^T Q_i Z_{ik})^2][X_{ik} - X_i^T W_i Z_i Q_i Z_{ik}]. \quad (32)$$

Multiplied by $\frac{-1}{48}$, Equation 32 is the derivative of the second last term in M_i.

For the last term in M_i,

$$\frac{\partial}{\partial \beta^T} k_i^T Q_i k_i = 2k_i^T Q_i \left\{ \left(\sum_j^{n_i} g_{ij} Z_{ij} Z_{ij}^T Q_i Z_{ij} \right) \left(X_{ij}^T - Z_{ij}^T Q_i Z_i^T W_i X_i \right) \right.$$

$$\left. - \left[\sum_j^{n_i} a_{ij} (Z_{ij}^T Q_i \otimes Z_{ij} Z_{ij}^T Q_i) \right] \left[\sum_k^{n_i} a_{ik} (Z_{ik} \otimes Z_{ik}) \right] \left[X_{ik}^T - Z_{ik}^T Q_i Z_i^T W_i X_i \right] \right\}$$

$$- (k_i^T \otimes k_i^T) \left[\sum_k^{n_i} a_{ik} (Q_i Z_{ik} \otimes Q_i Z_{ik}) \right] \left[X_{ik}^T - Z_{ik}^T Q_i Z_i^T W_i X_i \right].$$

$$= 2k_i^T Q_i \left(\sum_j^{n_i} g_{ij} Z_{ij} Z_{ij}^T Q_i Z_{ij} \right) \left(X_{ij}^T - Z_{ij}^T Q_i Z_i^T W_i X_i \right)$$

$$- 2k_i^T Q_i \left[\sum_j^{n_i} \sum_k^{n_i} a_{ij} a_{ik} Z_{ij} (Z_{ij}^T Q_i Z_{ik})^2 \right] \left[X_{ik}^T - Z_{ik}^T Q_i Z_i^T W_i X_i \right]$$

$$- \left[\sum_k^{n_i} a_{ik} (k_i^T Q_i Z_{ik})^2 \right] \left[X_{ik}^T - Z_{ik}^T Q_i Z_i^T W_i X_i \right]. \qquad (33)$$

The terms of Equation 33 that involve $2k_i^T Q_i$ are obtained by taking derivative of the k_i vectors on both sides of Q_i. To take derivative of $k_i (k_i = \sum_k^{n_i} a_{ik} B_{ik} Z_{ik})$, first use Equation F8 to get $2k_i^T Q_i$ in front of the derivative of k_i. Then take derivative of a_{ij} by using Equation 20, and derivative of Q_i inside k_i by using Equations F8, F9, 24 and F3. Finally, take derivative of Q_i between k_i^T and k_i, using Equations F8, F9 29 and F3. Thus, the transpose of Equation 33 is, since $k_i^T Q_i Z_{ij}$ is a scalar,

$$\frac{\partial}{\partial \beta} k_i^T Q_i k_i = 2 \left(\sum_j^{n_i} g_{ij} k_i^T Q_i Z_{ij} Z_{ij}^T Q_i Z_{ij} \right) \left(X_{ij} - X_i^T W_i Z_i Q_i Z_{ij} \right)$$

$$-2\left[\sum_{j}^{n_i}\sum_{k}^{n_i} a_{ij}a_{ik}k_i^T Q_i Z_{ij}(Z_{ij}^T Q_i Z_{ik})^2\right]\left[X_{ik} - X_i^T W_i Z_i Q_i Z_{ik}\right]$$

$$-\left[\sum_{k}^{n_i} a_{ik}(k_j^T Q_i Z_{ik})^2\right]\left[X_{ik} - X_i^T W_i Z_i Q_i Z_{ik}\right]. \qquad (34)$$

Multiplied by $\frac{15}{72}$, Equation 34 is the last term in the score function of the fixed effects β_i.

Score Function of the Variance-Covariance Components

This section will find the score function of the variance-covariance matrix of the random effects. It will prove that

$$S_{D_i} = \frac{-1}{2} vec(D^{-1}) + \frac{1}{2} vec(D^{-1} Q_i D^{-1}) - \frac{1}{2} \sum_j^{n_i} a_{ij} vec(D^{-1} Q_i Z_{ij} Z_{ij}^T Q_i Z_{ij} (y_i^* - \eta_i)^T W_i Z_i)$$

$$+ \frac{1}{2} vec(D^{-1} \tilde{b}_i \tilde{b}_i^T D^{-1})$$

$$+ \frac{1}{M_i} \left\{ -\frac{1}{8} \sum_j^{n_i} h_{ij} B_{ij}^2 \, vec(D^{-1} Q_i Z_{ij} (y_i^* - \eta_i)^T W_i Z_i) - \frac{1}{4} \sum_j^{n_i} g_{ij} B_{ij} vec(D^{-1} Q_i Z_{ij} Z_{ij}^T Q_i D^{-1}) \right.$$

$$+ \frac{1}{4} \sum_j^{n_i} \sum_k^{n_i} a_{ik} g_{ij} B_{ij} (Z_{ij}^T Q_i Z_{ik})^2 vec(D^{-1} Q_i Z_{ik} (y_i^* - \eta_i)^T W_i Z_i)$$

$$- \frac{1}{48} \sum_j^{n_i} p_{ij} B_{ij}^3 \, vec(D^{-1} Q_i Z_{ij} (y_i^* - \eta_i)^T W_i Z_i) - \frac{1}{16} \sum_j^{n_i} f_{ij} B_{ij}^2 vec(D^{-1} Q_i Z_{ij} Z_{ij}^T Q_i D^{-1})$$

$$+ \frac{1}{16} \sum_j^{n_i} \sum_k^{n_i} a_{ik} f_{ij} B_{ij}^2 (Z_{ij}^T Q_i Z_{ik})^2 vec(D^{-1} Q_i Z_{ik} (y_i^* - \eta_i)^T W_i Z_i)$$

$$+ \frac{15}{36} \sum_j^{n_i} g_{ij} (k_i^T Q_i Z_{ij}) (Z_{ij}^T Q_i Z_{ij}) vec(D^{-1} Q_i Z_{ij} (y_i^* - \eta_i)^T W_i Z_i)$$

$$+ \frac{15}{72} vec(D^{-1} Q_i k_i k_i^T Q_i D^{-1}) - \frac{15}{72} \sum_k^{n_i} a_{ik} (k_i^T Q_i Z_{ik})^2 vec(D^{-1} Q_i Z_{ik} (y_i^* - \eta_i)^T W_i Z_i)$$

$$+ \frac{15}{36} \sum_j^{n_i} a_{ij} (k_i^T Q_i Z_{ij}) vec(D^{-1} Q_i Z_{ij} Z_{ij}^T Q_i D^{-1})$$

$$\left. - \frac{15}{36} \sum_j^{n_i} \sum_k^{n_i} a_{ik} a_{ij} (k_{ij}^T Q_i Z_{ij}) (Z_{ij}^T Q_i Z_{ik})^2 vec(D^{-1} Q_i Z_{ik} (y_i^* - \eta_i)^T W_i Z_i) \right\}$$

Proof

By applying Equations F6 and F7,

$$\frac{\partial}{\partial(vecD)^T}\left(\frac{-1}{2}\log|D|\right) = \frac{-1}{2}\left(vec(D^{-1})\right)^T. \tag{35}$$

Then, take the transpose function off Equation 35, which is the first term in the score function.

For $\frac{1}{2}\log|Q_i|$, using Equation F6, F7, F9, 22 and F10 we have

$$\frac{\partial}{\partial(vecD)^T}\left(\frac{1}{2}\log|Q_i|\right) = \frac{\partial}{\partial(vecD)^T}\left(\frac{-1}{2}\log|D^{-1} + Z_i^T W_i Z_i|\right)$$

$$= \frac{-1}{2} tr(Q_i(\frac{\partial vec(D^{-1} + Z_i^T W_i Z_i)}{\partial(vecD)^T})$$

$$= \frac{1}{2}(vecD^{-1}Q_iD^{-1})^T - \frac{1}{2}(vecQ_i)^T\sum_j^{n_i} a_{ij}(Z_{ij}(y_i^* - \eta_i)^T W_i Z_i \otimes Z_{ij}Z_{ij}^T Q_i D^{-1})$$

$$= \frac{1}{2}[vec(D^{-1}Q_iD^{-1})]^T - \frac{1}{2}\sum_j^{n_i} a_{ij}[vec(D^{-1}Q_iZ_{ij}Z_{ij}^T Q_i Z_{ij}(y_i^* - \eta_i)^T W_i Z_i)]^T, \tag{36}$$

where

$$\frac{\partial vec(D^{-1} + Z_i^T W_i Z_i)}{\partial(vecD)^T} = [-(D^{-1}\otimes D^{-1}) + (Z_i^T \otimes Z_i^T)A_i((y_i^* - \eta_i)^T W_i Z_i \otimes Z_i Q_i D^{-1})]. \tag{37}$$

Take the transpose off Equation 36 to get the second and third terms of the score function S_{D_i}.

$$\frac{\partial \tilde{l}_i}{\partial(vecD)^T} = [vec(D^{-1}Q_iZ_i^T W_i(y_i^* - \eta_i)(y_i^* - \eta_i)^T W_i Z_i)]^T, \tag{38}$$

using Equations 22 and F10.

$$\frac{\partial}{\partial(vecD)^T} vec(\frac{-1}{2}\tilde{b}_i^T D^{-1}\tilde{b}_i) = -\frac{1}{2}(\tilde{b}_i^T \otimes \tilde{b}_i^T)\frac{\partial vec(D^{-1})}{\partial(vecD)^T} - \tilde{b}_i^T D^{-1}\frac{\partial \tilde{b}_i}{\partial(vecD)^T}$$

$$= \frac{1}{2}[vec(D^{-1}\tilde{b}_i\ \tilde{b}_i^T D^{-1})]^T - [vec(D^{-1}Q_i D^{-1}\tilde{b}_i(y_i^* - \eta_i)^T W_i Z_i)]^T, \qquad (39)$$

using Equations F8, F9, F11 and 21.

In adding up Equations 38 and 39, however, because of the fact that

$\tilde{b}_i = DZ_i^T W_i(y_i^* - \eta_i)$,

$[vec(D^{-1}Q_i Z_i^T W_i(y_i^* - \eta_i)(y_i^* - \eta_i)^T W_i Z_i)]^T - [vec(D^{-1}Q_i D^{-1}\tilde{b}_i(y_i^* - \eta_i)^T W_i Z_i)]^T = 0$,

with only $\frac{1}{2}[vec(D^{-1}\tilde{b}_i\ \tilde{b}_i^T D^{-1})]^T$ left, the transpose of which is the fourth term in the score function.

For the derivative of the first term in M_i

$$\frac{\partial}{\partial(vecD)^T}\sum_j^{n_i} g_{ij}B_{ij}^2 = \sum_j^{n_i} h_{ij}B_{ij}^2 Z_{ij}^T\,[(y_i^* - \eta_i)^T W_i Z_i \otimes Q_i D^{-1}]$$

$$-2\sum_j^{n_i} g_{ij}B_{ij}(Z_{ij}^T \otimes Z_{ij}^T)(Q_i \otimes Q_i)\frac{\partial vec(D^{-1} + Z_i^T W_i Z_i)}{\partial(vecD)^T}$$

$$= \sum_j^{n_i} h_{ij}B_{ij}^2\,[vec(D^{-1}Q_i Z_{ij}(y_i^* - \eta_i)^T W_i Z_i)]^T + 2\sum_j^{n_i} g_{ij}B_{ij}[vec(D^{-1}Q_i Z_{ij}Z_{ij}^T Q_i D^{-1})]^T$$

$$-2\sum_j^{n_i}\sum_k^{n_i} a_{ik}g_{ij}B_{ij}(Z_{ij}^T Q_i Z_{ik})^2[vec(D^{-1}Q_i Z_{ik}(y_i^* - \eta_i)^T W_i Z_i)]^T, \qquad (40)$$

using Equations 22, F8, F9, 37 and F10. The transpose of Equation 40 multiplied by $\frac{-1}{8}$ is the derivative of $\frac{-1}{8}\sum_j^{n_i} g_{ij}B_{ij}^2$ with respect to *vecD*.

Similarly,

$$\frac{\partial}{\partial(vecD)}\sum_{j}^{n_i} f_{ij}B_{ij}^3 = \sum_{j}^{n_i} p_{ij}B_{ij}^3\ vec(D^{-1}Q_iZ_{ij}(y_i^* - \eta_i)^T W_iZ_i)$$

$$+3\sum_{j}^{n_i} f_{ij}B_{ij}^2 vec(D^{-1}Q_iZ_{ij}Z_{ij}^TQ_iD^{-1})$$

$$-3\sum_{j}^{n_i}\sum_{k}^{n_i} a_{ik}f_{ij}B_{ij}^2(Z_{ij}^TQ_iZ_{ik})^2 vec(D^{-1}Q_iZ_{ik}(y_i^* - \eta_i)^T W_iZ_i). \qquad (41)$$

Equation 41 multiplied by $\frac{-1}{48}$ is the derivative of $\frac{-1}{48}\sum_{j}^{n_i} f_{ij}B_{ij}^T$ with respect to *vecD*.

The derivative of the last term in M_i is

$$\frac{\partial}{\partial(vecD)^T}k_i^TQ_ik_i = 2k_i^TQ_i\left\{\sum_{j}^{n_i} g_{ij}Z_{ij}Z_{ij}^TQ_iZ_{ij}Z_{ij}^T[(y_i^* - \eta_i)^TW_iZ_i \otimes Q_iD^{-1}]\right.$$

$$\left.-\sum_{j}^{n_i} a_{ij}(Z_{ij}^T \otimes Z_{ij}Z_{ij}^T)(Q_i \otimes Q_i)\frac{\partial vec(D^{-1} + Z_i^TW_iZ_i)}{\partial(vecD)^T}\right\}$$

$$-(k_i^T \otimes k_i^T)(Q_i \otimes Q_i)\frac{\partial vec(D^{-1} + Z_i^TW_iZ_i)}{\partial(vecD)^T}$$

$$= 2\sum_{j}^{n_i} g_{ij}(k_i^TQ_iZ_{ij})B_{ij}[vec(D^{-1}Q_iZ_{ij}(y_i^* - \eta_i)^TW_iZ_i)]^T$$

$$+2\sum_{j}^{n_i} a_{ij}(k_i^TQ_iZ_{ij})[vec(D^{-1}Q_iZ_{ij}Z_{ij}^TQ_iD^{-1})]^T$$

$$-2\sum_{j}^{n_i}\sum_{k}^{n_i} a_{ik}a_{ij}(k_{ij}^TQ_iZ_{ij})(Z_{ij}^TQ_iZ_{ik})^2[vec(D^{-1}Q_iZ_{ik}(y_i^* - \eta_i)^TW_iZ_i)]^T$$

$$+[vec(D^{-1}Q_ik_ik_i^TQ_iD^{-1})]^T - \sum_{k}^{n_i} a_{ik}(k_i^TQ_iZ_{ik})^2[vec(D^{-1}Q_iZ_{ik}(y_i^* - \eta_i)^TW_iZ_i)]^T. \qquad (42)$$

Again, we take derivative by using Equation F8, F9, F3, 22, 37 and F11.

The transpose of Equation 42 multiplied by $\frac{15}{72}$ is the derivative of the last term in M_1 with respect to *vecD*.

Chapter 4

AN ILLUSTRATIVE EXAMPLE

Introduction

This chapter presents an analysis on the data set of 1988 National Survey of Primary Education in Thailand (Thailand data). The analysis serves mainly as an example of the use of the multilevel Bernoulli model. It will also explore the differences and similarities among the four methods, namely, the first order Taylor expansion of the conditional expectation of the response (PQL), the second order Taylor expansion (PQL2), the sixth-order Laplace approximation to the log-likelihood (Laplace6) and Gauss-Hermite Quadrature approximation to the log-likelihood (Gauss). In addition, the differences produced by Gauss in using different numbers of quadrature points will also be of interest in this chapter.

Thailand Data

The Thailand data (USAID contract DPE-5824-A00-5076-00) were collected in 1988 by a research team from College of Education, Michigan State University, and Royal Thai Government, Office of the National Educational System. Information gathered includes survey and case studies. The purpose of the project was to "provide reliable data related to outcomes and costs of education and to allow study of policy

alternatives to improve the quality of primary education." (Taoklam et. al., 1992; See also Raudenbush and Bhumirat, 1992 ; Raudenbush, Bhumirat and Kamali, 1992)

The survey employed a multi-stage stratified sampling design. Samples were drawn at levels of schools and individuals. First, 405 schools were selected randomly within provinces. Then, one sixth grade class per school that had engaged in the national assessment project was selected at random from selected schools. At the individual level, samples were drawn from four population groups: principals, teachers, parents, and students. Student data are the interest of the current study. Information about schools where the classes were drawn was also collected. Altogether, 405 schools were sampled, within which data on 8582 pupils were collected. However, after deleting missing information of schools, data of 376 schools with 7877 students were used for the current analysis.

Before the survey began, Thailand had launched various programs since 1980 to improve the quality of education. These included a pre-primary education program, a national testing program, and various staff development programs for principals and teachers. The purpose of pre-primary education program was to improve each student's readiness for schooling. At the same time the government tried to promote the quality of administration and classroom teaching through staff development programs, and hold educators accountable for student learning through national testing programs. By requiring students to demonstrate basic skills before they can advance to the next grade, the country strove to ensure the quality of the product of school education --- student learning. It would be expected that the programs did help elevate educational efficiency.

Therefore, the research question here is whether students receiving pre-primary education, controlling for student and school background, had a smaller probability of repeating a grade.

However, an important variable that needs to be taken into account before making any claims about our focus of interest is socioeconomic status (SES). SES has always been found positively correlated with student achievement. I suspect this would also be true in Thailand. Whether students have adequate nutrition, especially breakfast, is an interesting variable. Students that did not have breakfast every day either came from poor families that could not afford breakfast every day, or had parents who did not pay too much attention to the children. Either way, not having breakfast interferes with students' concentration on learning, which might increase the probability of repeating grades. Finally, whether a student spoke central Thai dialect could also affect his or her probability of repeating a grade, since central Thai was the language used in class. If the student could not speak central Thai he or she would have difficulty understanding the instruction, which would increase the probability of repeating a grade. In addition, student gender is also an interesting covariate to put in, in order to see if girls do differently from boys in grade repeating. Therefore, student-level variables include:

response variable --- whether the student repeated grade(s) (REP1, 1 = yes, 0 = no);

variable of interest --- whether the student received pre-primary education (PPED1D, 1 = yes, 0 = no); and concomitant variables, which are

the student's gender (DSSEX, 1 = male, 0 = female);

student's family socio-economic background (SESC) (grand mean centered);

whether the student had breakfast every day (BRF1, 1 = yes, 0 = no); and

whether the student spoke dialects other than central Thai (DIALCT1, 0 = yes, 1 = no).

On the other hand, school environments may also affect student learning. Schools located in urban and affluent areas would have a larger enrollment and more resources than schools in rural and poor areas. Students from poor families who attend a big school might then have a better chance in education than students going to a poorer school. The average SES of students in a school is also a good indicator of the resources in a school. The average number of textbooks per student had in one school is a direct indicator for the instructional resources to which students have access. Without sufficient textbooks, it would be very difficult for students to learn. Thus, school information of interest includes:

natural log of school enrollment, grand mean centered (L_ENRC);

the average of students' SES, grand mean centered (MSESC); and

the average of number of books per student, grand mean centered (MTXBKC).

Results

After some preliminary runs, I found that the regression coefficients for variables in the first level either did not have significant amount of variance themselves, such as PPEDID (pre-primary education), DSSEX (student gender), or their variation could be explained by level-2 variables, such as SESC (student family SES), DIALCT1 (student

spoke dialect), and BREF1 (breakfast). Therefore, I decided to have a univariate random effect model. The first-level model for the data set is

$$\eta_{ij} = \alpha_{0i} + \alpha_{1i}*(SESC)_{ij} + \alpha_{2i}*(DSSEX)_{ij} + \alpha_{3i}*(DIALCT1)_{ij}$$
$$+ \alpha_{4i}*(BREF1)_{ij} + \alpha_{5i}*(PPEDID)_{ij},$$

while in level-2,

$$\alpha_{0i} = \beta_{00} + \beta_{01}*(L\_ENRC)_i + \beta_{02}*(MSESC)_i + b_i$$
$$\alpha_{1i} = \beta_{10} + \beta_{11}*(MSESC)_i$$
$$\alpha_{2i} = \beta_{20}$$
$$\alpha_{3i} = \beta_{30} + \beta_{31}*(MTXTBKC)_i$$
$$\alpha_{4i} = \beta_{40} + \beta_{41}*(L\_ENRC)_i$$
$$\alpha_{5i} = \beta_{50},$$

where $b_i \sim N(0,\ D_{00})$.

Table 1 - Estimates of Thailand Data

	PQL	PQL2	Gauss-10	Gauss-20	Gauss-30	Gauss-40	Laplace6
β_{00}	-2.0137 (.1409)	-2.2166 (.1524)	-2.2353 (.1429)	-2.2009 (.1420)	-2.1990 (.1421)	-2.1998 (.1421)	-2.1940 (.1421)
β_{01}	-.4031 (.1600)	-.4136 (.1781)	-.4614 (.1933)	-.4095 (.1909)	-.4159 (.1915)	-.4156 (.1914)	-.4147 (.1914)
β_{02}	-.6794 (.2606)	-.7889 (.2958)	-.7884 (.3055)	-.7845 (.3079)	-.7814 (.3079)	-.7809 (.3079)	-.7753 (.3076)
β_{10}	-.4971 (.1003)	-.5223 (.1056)	-.5325 (.1027)	-.5220 (.1035)	-.5220 (.1034)	-.5220 (.1034)	-.5223 (.1034)
β_{11}	.4657 (.1408)	.5003 (.1562)	.5321 (.1658)	.4962 (.1651)	.4976 (.1651)	.4972 (.1651)	.4978 (.1651)
β_{20}	.5549 (.0728)	.5819 (.0764)	.5840 (.0710)	.5825 (.0704)	.5825 .0704)	.5825 (.0704)	.5827 (.0704)
β_{30}	.3005 (.1262)	.3358 (.1384)	.3658 (.1235)	.3255 (.1304)	.3336 (.1300)	.3336 (.1301)	.3319 (.1300)
β_{31}	-.1012 (.0593)	-.1112 (.0655)	-.1513 (.0671)	-.1052 (.0781)	-.1114 (.0776)	-.1109 (.0776)	-.1104 (.0777)
β_{40}	-.4154 (.1032)	-.4327 (.1081)	-.4214 (.1041)	-.4354 (.1026)	-.4335 (.1028)	-.4340 (.1028)	-.4337 (.1028)
β_{41}	.2739 (.1355)	.2907 (.1461)	.2905 (.1447)	.2911 (.1440)	.2910 .1440)	.2910 (.1440)	.2910 (.1440)
β_{50}	-.4146 (.0947)	-.4501 (.1007)	-.4555 (.0993)	-.4462 (.0994)	-.4489 (.0994)	-.4482 (.0994)	-.4478 (.0994)
D_{00}	1.0703 (.1187)	1.444 (.1543)	1.473	1.383	1.390	1.388	1.3771 (.1830)

** numbers inside the parenthesis are standard errors

Table 1 gives the estimates of the fixed effects and the variance by Laplace6, Gauss-10, Gauss-20, Gauss-30 (30 quadrature points) and Gauss-40 (40 quadrature points), PQL, and PQL2. Although the methods give different estimates for the parameters (Table 1), they agree on the .05 significance level for all estimates, except for β_{31} by Gauss-10. In fact, the independent variables, except β_{31}, are all very powerful predictors for grade repetition. Especially the school level variables, L_ENRC (log-

enrollment) and MSESC (school mean SES) not only have great impact on the intercept, β_{00}, but also help predict, respectively, the impact of whether the student had breakfast every day (BREF1) and that of the student's personal SES background (SESC). However, their impacts on BREF1 and SESC are smaller than, and in opposite directions to, those on the intercept. That is, while students who did not have breakfast every day (BREF1) and came from family with low SES (SESC) had an increased risk of repeating grades, the effects of these adverse personal background are weakened if they attended big schools (high school enrollment) with higher school mean SES. In other words, an affluent school environment provides a cushion for students from poverty, helping prevent them from failing in school.

Pre-primary education, our focus of interest, also helped prevent a student from repeating grades. According to the preliminary runs, there is not much variation in its effect. Therefore, the effect of pre-primary education on grade repetition was pretty stable across different schools. On the other hand, students speaking central Thai also tended to have advantage in their learning. Having textbooks helped reduce the disadvantage of speaking dialects by about one third of the effects of speaking dialects other than central Thai. This makes sense since students could learn little by little on their own if they had textbooks at hand. However, the effect is not significant at .05 level. Its p-value is around .15. Finally, girls did seem to learn better, in primary school level, than boys. Holding all other variables at the average, a boy had a higher logit of around .58 of repeating grades than girls.

The comparison among the estimates by different methods is another interesting issue. As shown in the table, a lot of the differences between Gauss-10 and Gauss-20 are in the second decimal place. Gauss-20 and Gauss-30 differ in the third decimal place. Gauss-30 and Gauss-40 do not differ too much, only at the fourth decimal place. Some of Laplace6 results differ from Gauss-40 in the third place and some in the fourth place. Laplace6 results are generally closer to Gauss-30 and Gauss-40 than Gauss-10 and Gauss-20. PQL2 and PQL are further away. PQL consistently gives estimates that are smaller in absolute values. PQL2 results are actually pretty close to those of Gauss's with larger numbers of points, but they are not as close as those of Laplace6.

Chapter 5

EVALUATION WITH SIMULATED DATA

Introduction

This chapter compares the 4 methods, Laplace6, Gauss-Hermite Quadrature (Gauss) (Hedeker and Gibbons, 1994; 1996), PQL (Raudenbush, 1993) and PQL2 (Goldstein and Rasbash, 1996) by analyzing data sets simulated under 8 different models. The comparison will be in terms of 1) the unbiasedness of the estimates ($(\hat{\beta}\ ,\ vec\hat{D}) = \hat{\theta}$) across data sets under the same model; 2) the mean squared errors of the estimates; 3) average of standard errors from outputs ($\theta_{S\hat{E}}$); 4) standard deviation of the estimates across data sets ($SD(\hat{\theta})$); and 5) the relative efficiency of Laplace6 to the other methods.

Eight different models were used to simulate data sets. The first six models (Models 1 to 6) were univariate random effect models that had a wide range (.52, .2, .1) of the average conditional expectations of the response y_{ij} given $b_i = 0$ ($\bar{\mu}_{ij}^{(0)} = E(E(y_{ij}|b_i = 0))$) and two different values for the random effect variance, namely, 1, .25. The data sets were generated by Yang (1994). The purpose of the use of the six different models was to investigate whether the methods performed differently

depending on parameter values. Presumably, models with $\bar{\mu}_{ij}^{(0)}$ close to .5 will be the easiest for all methods, because of the symmetry of the data sets. As $\bar{\mu}_{ij}^{(0)}$ becomes smaller, the estimation task will become more difficult. However, while PQL was already known to have a large negative bias for larger variances, the performance of the other methods for large variances was of interest.

Two bivariate random effects models (Models 7 and 8) were constructed to assess the four methods with dependent random effects, in two ways. Model 7 explored the performances of the methods under severe conditions with small $\bar{\mu}_{ij}^{(0)}$=.143 and extreme values (1.625, .25) for the variances with a small covariance (.1). The interest of Model 8 lay in the wish to inspect the consistency property of the maximum likelihood estimates produced by Laplace6. The investigation was launched by comparing estimates under the same model but with two different cluster sizes in the second level, the first set being 10 times smaller than the second set. The property of consistency would be revealed if there is little bias in the estimates and the variances of estimates become smaller as the sample size increases.

The basic structure of the data sets followed Rodriguez and Goldman (1995). In the first level, we had $\eta_{ij} = \log[\mu_{ij} / (1-\mu_{ij})] = \alpha_{0i} + (childc)_{ij} * \alpha_{1i}$. In the second level, $\alpha_{oi} = \beta_{00} + (commuc)_i * \beta_{01} + b_{0i}$ with $b_{0i} \sim N(0, D_{00})$. Here $\alpha_{1i} = \beta_{10}$ was fixed for the first six univariate random effects models. For bivariate random effects (Models 7 and 8), $\alpha_{1i} = \beta_{10} + b_{1i}$ was random with $b_{1i} \sim N(0, D_{11})$, and $\text{cov}(b_{0i}, b_{1i}) = D_{01}$. The values of β_{01} and β_{10} were both set to 1. The values for β_{00} were manipulated in order

to get different values for $\bar{\mu}_{ij}^{(0)}$. The level-1 covariate, childc, was sampled from a normal distribution with mean .0955621, and variance .0676, while the level-2 predictor, commuc, was sampled from a normal distribution with mean -.6857591 and variance .2304. However, in bivariate random effects models, the means of both covariates remain unchanged while their variances were both changed to 1. There was no missing value in any of these models.

Univariate Random Effect Data Sets (Models 1 - 6)

The six univariate random effect models used 3 different values for the average conditional probabilities, $\bar{\mu}_{ij}^{(0)}$, namely, .52, .2, and .1, and 2 different values of variance D_{00}, 1, and .25, where 1 is usually supposed to be large and .25 pretty small. The values of β_{00} were .6653, -.7961, and -1.62 for $\bar{\mu}_{ij}^{(0)}$ to be .52, .2 and .1, respectively. Each data set had 16 observations for each cluster in the first level, and 161 clusters in the second level. For each combination of the parameters, 50 data sets were generated. Model 1 had $\bar{\mu}_{ij}^{(0)}$=.52, D_{00}= 1 , while Model 2 had the same value for $\bar{\mu}_{ij}^{(0)}$, but a smaller variance, D_{00}=.25. Model 3 had $\bar{\mu}_{ij}^{(0)}$ =.2 and D_{00} = 1, whereas Model 4 differed from Model 3 by a smaller variance, D_{00}=.25. Similarly, under Model 5, $\bar{\mu}_{ij}^{(0)}$ = .1, D_{00} = 1; under Model 6, $\bar{\mu}_{ij}^{(0)}$ =.1, τ_{00}=.25.

Gauss results were computed using 10 quadrature points (Gauss-10). The results were obtained from Yosef (1997). Ten points were specified because, according to the MIXOR manual(1993), 8 to 10 points would produce satisfactory results for univariate data sets, whereas fewer points could be specified for higher dimensional data sets. PQL

was not compared here since Yosef (1997) has found it to consistently underestimate the fixed effects and the variance components, in accordance with previous results (e.g., Goodman and Rodriguez, 1995; Breslow and Clayton, 1993).

Results of Model 1

Table 2 - Averages and Mean Squared Errors of Model 1

	Laplace6		Gauss-10		PQL2	
	average	mse	average	mse	average	mse
D_{00}=1	1.0135	0.0294	1.0142	.0300	1.0361	0.0336
β_{00}=.665267	0.6679	0.0251	0.6677	.0251	0.6750	0.0258
β_{01}=1	0.9812	0.0430	0.9835	.0429	0.9913	0.0440
β_{10}=1	0.9891	0.0400	0.9901	.0401	0.9944	0.0405

The clearest pattern in Table 2 is that, under Model 1, the averages and mean squared errors of the three methods were very close to each other, although PQL2 consistently had a slightly larger mean squared errors than the other two. The biases of the three methods were small, and the directions of biases for the parameters were the same too. Another clear pattern is that PQL2 always gave the largest estimate for all the parameters, whether the bias of the three methods was negative or positive for a particular parameter. The amount of positive bias of PQL2 for the variance was 3.6% of the parameter, which seemed to be a little too large compared to that of the other two methods.

Table 3 - Averages of S. E.'s and S. D.'s of Estimates of Model 1

	Laplace6		Gauss-10		PQL2	
θ	θ_{SE}	$SD(\theta)$	θ_{SE}	$SD(\theta)$	θ_{SE}	$SD(\theta)$
D_{00}=1	.1759	.1725	.1740	.1742	NA	.1816
β_{00}=.665267	.1698	.1599	.1683	.1602	.1648	.1621
β_{01}=1	.2014	.2085	.1999	.2087	.1953	.2116
β_{10}=1	.1789	.2018	.1788	.2021	.1749	.2032

The standard error of D_{00} was not available in the PQL2 program. The averages of the standard errors (θ_{SE}) were the average amounts of uncertainty the methods predicted for the estimates. The standard deviations of the estimates indicated the real amounts of uncertainty in the estimation. The discrepancy between the prediction and the reality gathered from the 50 data sets was the largest for all three methods for β_{10}, for which all three methods underestimated the variability; and smallest for D_{00} by Laplace6 and Gauss-10. The differences between the averages of the standard errors and the standard deviations of the estimates were the smallest for Gauss-10.

Results of Model 2

Table 4 - Averages and Mean Squared Errors of Model 2

	Laplace6		Gauss-10		PQL2	
	average	mse	average	mse	average	mse
D_{00}=.25	.2656	.0048	.2658	.0048	.2662	.0049
β_{00}=.665267	.6759	.0111	.6760	.0111	.6771	.0112
β_{01}=1	1.0123	.0158	1.0124	.0158	1.0141	.0159
β_{10}=1	1.0010	.0380	1.0011	.0380	1.0025	.0381

Model 2 was different from Model 1 only in the value of D_{00}. Again, the three methods were very similar in both biasedness and mean squared errors. The mean squared errors of Laplace6 and Gauss-10 were identical. With a smaller value of D_{00}, the three methods all had positive bias, although small again. The largest positive bias appeared for D_{00}, at about 6% of the parameter by all three methods.

Table 5 - Averages of S. E.'s and S. D.'s of Estimates of Model 2

	Laplace6		Gauss-10		PQL2	
θ	θ_{SE}	$SD(\theta)$	θ_{SE}	$SD(\theta)$	θ_{SE}	$SD(\theta)$
D_{00}=.25	.0684	.0683	.0685	.0684	NA	.0687
β_{00}=.665267	.1074	.1060	.1075	.1060	.1045	.1061
β_{01}=1	.1264	.1265	.1264	.1265	.1243	.1267
β_{10}=1	.1686	.1970	.1686	.1970	.1643	.1973

The similar values for the standard deviations of the averages under Model 2 in Table 5 were consistent with the close similarity of the mean squared errors. The

prediction of the uncertainty by the three methods were generally pretty close to the empirical results. All three methods underestimated the variability of β_{10} by the largest amount, as in Model 1.

Results of Model 3

Table 6 - Averages and Mean Squared Errors of Model 3

	Laplace6		Gauss-10		PQL2	
	average	mse	average	mse	average	mse
D_{00} =1	.9396	.0472	.9362	.0463	.9772	.0515
β_{00}=-.7960974	-.7794	.0196	-.7801	.0199	-.7836	.0199
β_{01}=1	1.0254	.0328	1.0261	.0330	1.0361	.0347
β_{10}=1	1.0322	.0364	1.0324	.0364	1.0356	.0369

For Model 3, Laplace6 results also followed closely those of Gauss-10. Contrary to the situation in Model 1, the three methods had negative bias for D_{00} and very small positive biases for the β's. The underestimation of PQL2 for D_{00} was around 2% of the parameter, while that by Laplace6 and Gauss-10 was much larger, around 6%. The biases for the β's by the three methods were very close to each other. However, the mean squared errors for PQL2 were all slightly larger than those for the other two.

Table 7 - Averages of S. E.'s and S. D.'s of Estimates of Model 3

	Laplace6		Gauss-10		PQL2	
θ	θ_{SE}	$SD(\theta)$	θ_{SE}	$SD(\theta)$	θ_{SE}	$SD(\theta)$
D_{00}=1	.1789	.2108	.1766	.2076	NA	.2276
β_{00}=-.7960974	.1633	.1404	.1625	.1415	.1613	.1418
β_{01}=1	.2029	.1810	.2018	.1817	.1994	.1845
β_{10}=1	.2077	.1899	.2076	.1900	.2017	.1908

A significant pattern of Table 7 is that for Model 3, all the three methods seemed to under-predict the variation of the estimates of D_{00}, and over-predict those of all the other parameters. The largest difference between prediction and empirical results occurred for D_{00}. Laplace6 had the largest discrepancy among the three for all parameters, over-predicting the variations of the three fixed effects; while PQL2 had the smallest discrepancy.

Results of Model 4

Table 8 - Averages and Mean Squared Errors of Model 4

	Laplace6		Gauss-10		PQL2	
	average	mse	average	mse	average	mse
D_{00}=.25	.2435	.0059	.2427	.0059	.2501	.0060
β_{00}=-.7960974	-.7854	.0077	-.7853	.0077	-.7873	.0077
β_{01}=1	1.0057	.0144	1.0057	.0144	1.0075	.0145
β_{10}=1	1.0060	.0498	1.006	.0498	1.0071	.0499

With a small value of D_{00}, the mean squared errors of the three methods were almost identical, as in Model 2. The biases of D_{00} by PQL2 were almost 0, while the

negative bias by Laplace6 and Gauss-10 was around 3% of the parameter. The β_{00}'s were pretty much unbiased by all three methods.

Table 9 - Averages of S. E.'s and S. D.'s of Estimates of Model 4

	Laplace6		Gauss-10		PQL2	
θ	θ_{SE}	$SD(\theta)$	θ_{SE}	$SD(\theta)$	θ_{SE}	$SD(\theta)$
D_{00}=.25	.0816	.0774	.0814	.0770	NA	.0782
β_{00}=-.7960974	.1123	.0878	.1122	.0878	.1099	.0880
β_{01}=1	.1437	.1212	.1436	.1212	.1392	.1215
β_{10}=1	.2014	.2254	.2014	.2254	.1957	.2256

All three methods tended to over-predict the variation of the estimates, except for β_{10}. The discrepancies between the predicted and empirical variation of the estimates for the three methods were very close, too, although PQL2 had a slightly smaller discrepancy than the other two methods; the discrepancies for Laplace6 and Gauss-10 were almost identical.

Results of Model 5

Table 10 - Averages and Mean Squared Errors of Model 5

	Laplace6		Gauss-10		PQL2	
	average	mse	average	mse	average	mse
D_{00}=1	.9742	.0601	.9720	.0580	1.0511	.0803
β_{00}=-1.62	-1.6122	.0318	-1.6138	.0322	-1.6306	.0335
β_{01}=1	.9994	.0580	1.0016	.0593	1.0117	.0592
β_{10}=1	.9990	.0499	.9981	.0495	1.0006	.0502

Again, the results of Laplace6 went together closely with Gauss-10 in Model 5. For D_{00}, the negative bias of the two methods were both around 2.5% of the parameter, while PQL2 had a positive bias of 5%. This was different from experiences with the above models, where PQL2 always had the same signs for biases (positive or negative) as the other two methods. On the other hand, all the β's by the three methods were almost unbiased. Gauss-10's mean squared error of D_{00} was a little smaller than that of Laplace6, while the β's of Laplace6 had smaller mean squared errors. PQL2 generally had the largest mean squared errors for all parameters, as before.

Table 11 - Averages of S. E.'s and S. D.'s of Estimates of Model 5

	Laplace6		Gauss-10		PQL2	
θ	θ_{SE}	$SD(\theta)$	θ_{SE}	$SD(\theta)$	θ_{SE}	$SD(\theta)$
D_{00}=1	.2221	.2462	.2184	.2417	NA	.2815
β_{00}=-1.62	.1811	.1800	.1807	.1811	.1789	.1845
β_{01}=1	.2317	.2432	.2310	.2460	.2291	.2456
β_{10}=1	.2485	.2256	.2483	.2248	.2481	.2262

The standard deviation of D_{00} of Model 5 for PQL2 in Table 11 was much larger than those of the other two methods. This contributed to the large value of its mean squared error in Table 9. The three methods under-predicted the variation of D_{00} and over-predicted that of β_{10}. The discrepancies for the other two β's were small. The discrepancies by Gauss-10 were smaller than those of Laplace6 and PQL2.

Results of Model 6

Table 12 - Averages and Mean Squared Errors of Model 6

	Laplace6 (49 obs.)		Gauss-10 (48 obs.)		PQL2 (50 obs.)	
	average	mse	average	mse	average	mse
D_{00}=.25	.2389	.0117	.2370	.0118	.2593	.0142
β_{00}=-1.62	-1.6139	.0123	-1.6119	.0124	-1.6214	.0123
β_{01}=1	.9995	.0259	1.0044	.0252	1.0063	.0266
β_{10}=1	.9933	.0765	.9883	.0770	1.0017	.0777

For Model 6, Laplace6 gave converged results for 49 out of the 50 data sets, Gauss-10, 48 , while PQL2 had no difficulty with any of the data sets, as was shown in Table 12. Laplace6 results were again very close to those of Gauss-10, both in averages

and mean squared errors. PQL2 seemed more unbiased than the other two but it gave larger mean squared errors for the parameters. The negative bias of D_{00} by Gauss-10 and Laplace6 were both around 5% of the parameters, whereas PQL2's negative bias was a little smaller, around 3.5%. The three methods' estimates for the β 's were almost unbiased.

Table 13 - Averages of S. E.'s and S. D.'s of Estimates of Model 6

	Laplace6		Gauss-10		PQL2	
	θ_{SE}	$SD(\theta)$	θ_{SE}	$SD(\theta)$	θ_{SE}	$SD(\theta)$
D_{00}=.25	.1053	.1089	.1051	.1091	NA	.1198
β_{00}=-1.62	.1301	.1119	.1299	.1123	.1257	.1121
β_{01}=1	.1652	.1625	.1649	.1605	.1642	.1648
β_{10}=1	.2568	.2794	.2575	.2802	.2506	.2816

The standard deviation of the estimates of PQL2 in Table 13 were again the largest. However, it gave the smallest standard errors of the estimates, as in the models discussed above. The underestimation of the variation was most severe for β_{10}. The discrepancies between the predicted and empirical variation for the other two β 's were the smallest by PQL2. Laplace6 and Gauss-10 were very similar in the errors of prediction for the variations of β_{10} and D_{00}; both were smaller than those by PQL2.

Relative Efficiencies Under Models 1 to 6

Table 14 - Laplace6 Relative Efficiency Under Models with D_{00}=1

	Model 1 ($\bar{\mu}_{ij}^{(0)}$=.52)		Model 3 ($\bar{\mu}_{ij}^{(0)}$=.2)		Model 5 ($\bar{\mu}_{ij}^{(0)}$=.1)	
	Gauss-10	PQL2	Gauss-10	PQL2	Gauss-10	PQL2
D_{00}	1.0204	1.1429	.9809	1.0911	1.0362	1.3361
β_{00}	1.0000	1.0279	1.0153	1.0153	1.0354	1.0535
β_{01}	.9977	1.0233	1.0061	1.0579	1.0224	1.0207
β_{10}	1.0025	1.0125	1.0000	1.1014	.9920	1.0060

Table 15 - Laplace6 Relative Efficiency Under Models with D_{00}=.25

	Model 2 ($\bar{\mu}_{ij}^{(0)}$=.52)		Model 4 ($\bar{\mu}_{ij}^{(0)}$=.2)		Model 6 ($\bar{\mu}_{ij}^{(0)}$=.1)	
	Gauss-10	PQL2	Gauss-10	PQL2	Gauss-10	PQL2
D_{00}	1.0000	1.0208	1.0000	1.0169	1.0085	1.2137
β_{00}	1.0000	1.0090	1.0000	1.0000	1.0081	1.0000
β_{01}	1.0000	1.0063	1.0000	1.0069	.9730	1.0270
β_{10}	1.0000	1.0026	1.0000	1.0020	1.0065	1.0000

Tables 14 and 15 give the efficiencies of Laplace6 relative to Gauss-10 and PQL2. The relative efficiency for D_{00}, say, of Laplace6 to Gauss-10 is the ratio of Gauss-10's mean squared error for D_{00} to that of Laplace6's. Therefore, Laplace6 has higher efficiency for D_{00} if the ratio is larger than one, and vice versa. From the two tables, the efficiencies of Laplace6 relative to Gauss-10 were mostly larger or equal to 1. The relative efficiencies of Laplace6 were slightly higher for larger D_{00}; for smaller D_{00}, the relative efficiencies were mostly 1. The only exception was that of β_{01} under Model 6, where Gauss-10 had a higher efficiency. The effect of larger variance and smaller

average conditional expectation on the loss of efficiencies relative to Laplace6 was even more apparent in PQL2. Laplace6 was more efficient than PQL2 for the fixed effects and the variance. Gauss-10 was also more efficient than PQL2.

In conclusion, the three methods performed reasonably well under the six models. However, the values of D_{00} and $\overline{\mu}_{ij}^{(0)}$ did have an impact on how precisely and accurately the three methods estimated the parameters. With a smaller D_{00}, the variation (standard deviations) of the β_{10} estimates by all three methods were larger than what they predicted (the average of the standard errors). On the other hand, when the value of $\overline{\mu}_{ij}^{(0)}$ became smaller (.1 and .2), the three methods tended to have negative biases for D_{00}. Moreover, with small values of $\overline{\mu}_{ij}^{(0)}$(.1, .2) and a large D_{00}, the variation of D_{00} was underestimated by all three methods.

The biases of all the three programs were generally very small in these univariate models. The three programs almost always went together in the direction of biases. PQL2 always gave the largest absolute value of the estimates. Because of the largest variation in estimates, its mean squared errors were usually the largest among the 3 methods too, although the difference was usually small. Laplace6 estimates were very close to those of Gauss-10 in terms of averages and mean squared errors. Even in the discrepancies in the prediction of variation, Laplace6 results were very similar to those of Gauss-10, although the discrepancies seemed smaller for Gauss-10 in more cases. The largest disagreement between the Laplace6 and Gauss-10 in terms of both the averages of the standard errors (θ_{SE}) and the standard deviations of the estimates ($SD(\theta)$) were in the

third decimal place, while the largest disagreement between either Laplace6 or Gauss-10 and PQL2 was in the second decimal place.

However, Laplace6 and PQL2 gave estimates up to the fourth decimal place. Gauss gave variance and covariance estimates only to the third decimal place, computed from the values of the Cholesky decomposition rounded up to the second decimal place. Therefore, because of rounding errors, the mean squared errors of the variance and covariance estimates from Gauss might look larger than they really were. The comparison of variance and covariance estimates thus might not be exact.

Bivariate Data Sets

Model 7 and its Results

Model 7 contained 100 data sets with parameters $\beta_{00} = -1.2$, $\mu_{ij}^{(0)} = .143$, $D_{00} = 1.625$, $D_{01} = .1$, $D_{11} = .25$, $\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1.625 & .1 \\ .1 & .25 \end{pmatrix}\right)$. Each data set contained 20 observations in the first level and 200 clusters in the second level.

Laplace6 was computed on an old UNIX machine, using the converged estimates of the parameters from PQL. The 100 data sets took Laplace6 altogether 2 hours to analyze. PQL2 was computed on the same UNIX, too, using similar amount of time for the 100 data sets. The second order Taylor expansion of the conditional likelihood was set to start at the second iteration while computing starting values based on PQL. According to experiences, starting the Taylor expansion earlier would cost only some more iterations but would not have effect on converged values. However, if the starting values happened to converge fast and the Taylor expansion was set in after that, the

resulting estimate would be just PQL. All the 3 programs (Laplace6, PQL, and PQL2) were run in UNIX.

Gauss-Hermite Quadrature method (using MIXOR package) was run on Pentium 233. Out of curiosity for the use of more quadrature points in the improvement of the accuracy of the estimation and also intrigued by the pattern in the analysis of Thailand data, where the results of Laplace6 seemed to be more similar to Gauss with more points than with fewer points, both Gauss-10 (Gauss-Hermite Quadrature using 10 points) and Gauss-20 (Gauss-Hermite Quadrature using 20 quadrature points) were used in analyzing the 100 data sets. The estimates from the 4 programs were compared in terms of their unbiasedness and mean squared errors. Gauss-20 used about 20 hours altogether for the 100 data sets, and Gauss-10 used 5 hours.

Although it was impossible to compare the time used by the 3 different methods, i.e., PQL, PQL2, and Laplace6, with Gauss, 6 data sets were randomly selected to use the same Pentium 233 to do the analysis. The time for the 6 data sets used for PQL plus Laplace6 ranged from 7 seconds to 20 seconds; for the same 6 data sets, the time used by Gauss-20 was around 12 minutes, while the time used by Gauss-10 was about 3 to 5 minutes. Thus it was very clear that Laplace6 was significantly more efficient in terms of computational time.

Gauss program produces variances and covariances from the Cholesky decomposition and gives standard errors for the decomposed terms only. It was impossible to find out the standard errors of the variance and covariance from what is available.

Table 16 - Averages of Estimates Under Model 7

	Laplace6	Gauss-10	Gauss-20	PQL	PQL2
D_{00}=1.625	1.6352	1.6532	1.6546	1.2752	1.7296
D_{01}=.1	.0960	.1003	.0995	.0538	.0864
D_{11}=.25	.2667	.2575	.2562	.1614	.2927
β_{00}=-1.2	-1.2007	-1.1977	-1.2045	-1.0904	-1.2179
β_{01}=1	1.0029	1.0081	1.0148	.9004	1.0231
β_{10}=1	.9975	0.9971	.9976	.9114	1.0050

The estimates by all programs were fairly normally distributed. PQL again consistently had negative bias for all the parameters. The bias of the variance components ranged from 22% (D_{00}) to 46% (D_{01}) of the parameters, while those of the β 's were around 9% of the parameters. PQL2 had the second highest bias, either positive or negative. It had positive bias for D_{00}, by about 6.5%, and D_{11} by 17%, but negative bias for D_{01}, by 14%. Even though the positive biases of its β estimates were only around 1%, they were still larger than the two Gauss's and Laplace6. The advantage of Gauss-20 over Gauss-10 was not very clear from the table, since the averages of the two were very similar. Laplace6 results were again close to those of the Gauss's. Laplace6 had more biases, both negative and positive, than the Gauss's for smaller values of the variance components, but had smaller positive bias for the large variance than the latter. For the β's, the three methods were pretty much unbiased.

Table 17 - Mean Squared Errors of Estimates Under Model 7

	Laplace6	Gauss-10	Gauss-20	PQL	PQL2
D_{00}=1.625	.0563	.0737	.0633	.1522	.0838
D_{01}=.1	.0108	.0115	.0120	.0080	.0143
D_{11}=.25	.0075	.0073	.0072	.0113	.0115
β_{00}=-1.2	.0190	.0231	.0196	.0271	.0203
β_{01}=1	.0164	.0193	.0175	.0236	.0178
β_{10}=1	.0051	.0051	.0053	.0116	.0055

The mean squared errors of Table 17 tell another different story. Laplace6 produced the smallest, or among the smallest, mean squared errors for the estimates among the five methods. However, considering that estimates of D_{01} and D_{11} by Laplace6 had relatively large amounts of bias but that their mean squared errors were either smaller or only .0001 larger than those of Gauss-20, the variation of the variance components by Laplace6 seemed to be much smaller than that of the Gauss's. The comparison of Gauss-10 with Gauss-20 was clearer in terms of mean squared errors. Gauss-20 most of the time had much smaller mean squared errors than Gauss-10, the values were closer to those of Laplace6 than those of Gauss-10, too. The mean squared errors of PQL were notably larger. The mean squared error for D_{01} was the smallest of its counterparts of the other three methods, although its underestimation from Table 3 was 46%. This again indicated that PQL performed well for small values of random effect (co)variance. On the other hand, considering that D_{01} = 0.1, and that the mean squared errors of D_{01} by the other methods were all larger than their respective mean squared error for D_{11}=0.25, it seemed that all the other 3 methods had a difficult time giving a

reasonable estimate for D_{01}. PQL2 produced the second largest mean squared errors for the variances and covariance of the random effects. However, its mean squared errors of the 3 β's were only a little larger than those of Gauss-20. PQL2 seemed to perform better for the fixed effects than for the variance components.

Table 18 - Laplace6 Relative Efficiency Under Model 7

	Gauss-10	Gauss-20	PQL	PQL2
D_{00}=1.625	1.3091	1.1243	2.7034	1.4885
D_{01}=.1	1.0648	1.1111	.7407	1.3241
D_{11}=.25	.9733	.9600	1.5066	1.5333
β_{00}=-1.2	1.2158	1.0316	1.4263	1.0684
β_{01}=1	1.1768	1.0671	1.4390	1.0854
β_{10}=1	1.0000	1.0392	2.2745	1.0784

The relative efficiencies of Laplace6 relative to all the other programs in Table 18 gave a clear picture of the comparison of the mean squared errors. Laplace6 was more efficient than all the other methods in general, except for D_{11}, for which Laplace6 had a positive bias of 6% of its value (Table 16). However, even though Laplace6 also had a 4% positive bias for D_{01}, it was still more efficient than Gauss-10 and Gauss-20, whose estimates were almost unbiased. The extra ten quadrature points in Gauss were a mixed blessing. It seemed that Gauss-20 was not as inefficient as Gauss-10 when the latter fell quite far behind Laplace6. However, at times, it was a little less efficient than Gauss-10 when the latter was only a little less efficient than Laplace6.

In Table 19, the standard deviations of the estimates by PQL again were the smallest among the five methods. PQL2 had the largest standard deviations of the estimates of D_{01}, D_{11} and β_{10}. Gauss-10 had the largest standard deviations of the estimates of β_{00}, β_{01} and D_{00}. The variation of Laplace6 estimates was the smallest among the four programs without considering PQL. The small amounts of variation in the estimates, in addition to small biases, contributed to the significantly smaller mean squared errors and higher efficiency. The variations of the estimates by Gauss-20 was not necessarily smaller than those of Gauss-10, although its variation for D_{00} was indeed much smaller.

Table 19 - Standard Deviation of the Estimates

	Laplace6	Gauss-10	Gauss-20	PQL	PQL2
D_{00}=1.625	.2383	.2714	.2510	.1737	.2713
D_{01}=.1	.1045	.1076	.1102	.0768	.1194
D_{11}=.25	.0853	.0857	.0853	.0593	.0988
β_{00}=-1.2	.1387	.1529	.1403	.1233	.1421
β_{01}=1	.1288	.1393	.1320	.1171	.1320
β_{10}=1	.0716	.0721	.0730	.0615	.0744

Table 20 - Averages of Standard Errors

	Laplace6	Gauss-10	Gauss-20	PQL	PQL2
D_{00}=1.625	.2684	.2563	.2688	.1831	.2420
D_{01}=.1	.1156	NA	NA	.0786	.1046
D_{11}=.25	.0956	NA	NA	.0662	.0894
β_{00}=-1.2	.1290	.1198	.1293	.1110	.1273
β_{01}=1	.1175	.1105	.1170	.0999	.1162
β_{10}=1	.0755	.0744	.0747	.0602	.0701

In reference to Tables 19 and 20, the averages of the standard errors produced by the methods were compared to their respective real standard deviations. PQL seemed to have the smallest discrepancies between the two tables for all parameters. PQL2 had the largest discrepancies. The variation of all its estimates were under-predicted by the standard errors it gave. Laplace6 over-predicted the variation of the variance components and of β_{10}, and under-predicted the other 2 β's. Gauss-10 and Gauss-20 had the same pattern as Laplace6 in estimating the variation of the fixed effects, but Gauss-10 underestimated that of D_{00} while Gauss-20 was in the opposite direction. Apart from

PQL, Gauss-20 seemed to have the smallest discrepancy between the theoretical standard errors for its estimates and their empirical standard deviation.

In summary, Laplace6 produced estimates that were very close to those of Gauss-20, both in averages and in mean squared errors. Their biases were reasonably small. Over all, Laplace6 had the smallest mean squared errors and the highest efficiency, thanks to its least variability across data sets. The discrepancy between the theoretical variation of the estimates and its empirical variation was the smallest in PQL. Gauss-20 has the second smallest discrepancy. The advantage of 20 quadrature points over 10 quadrature points was clear also in the parameters where Gauss-10 had substantially larger amounts of mean squared errors than Laplace6. For these parameters, the mean squared errors of Gauss-20 were much smaller. However, for parameters where Gauss-10 had only a little larger, or smaller, mean squared errors than Laplace6, Gauss-20 might do slightly worse than Gauss-10. This might suggest that the advantage of using a larger number of points for Gauss appears only where accurate estimation is difficult using a smaller number of points. However, with real data sets, it is impossible to decide on the accuracy of the estimates. The biases (in percentage of the parameters) of PQL2 under the current model were larger than under the univariate models. Its efficiencies of all estimates relative to Laplace6 decreased a lot in this bivariate model, too. As to the prediction of the variation of estimates, consistent with univariate models, PQL2 gave smaller estimates of the variation than the empirical results.

Model 8 and its Results

Model 8 contained 400 data sets, which had parameters, $\beta_{00} = -.508403$, $D_{00} = 2$, $D_{01} = .2$, $D_{11} = .75$,, $\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & .2 \\ .2 & .75 \end{pmatrix}\right)$. Each of the data sets had 20 observations for each cluster in the first level. In the second level, 200 of the 400 data sets had 200 clusters in each of them, while there were 2000 clusters in each of the latter 200 data sets.

The interest was to check whether Laplace6 estimates were consistent as cluster sizes increased by 10 times. That is, if Laplace6 estimates using the latter 200 data sets have smaller biases and 10 times smaller variance than the former 200 do, Laplace6 method is considered consistent. We would also be interested in the percentages the estimates fell beyond the 95% confidence interval of the true parameter (i.e., true value ± 1.96 standard deviations), using the empirical standard deviations. This would give a sense of how well the estimation was. Besides, if the estimates were normally distributed then the theoretical probability of observations falling beyond the 95% confidence interval (i.e., average ± 1.96 standard deviations) is .05.

Therefore, statistics produced for each set of estimates using data sets of 200 clusters would be contrasted with those using data sets of 2000 clusters in the second level. These included averages, biases, variances, the empirical probabilities of observations falling out of the range of the true value ± 1.96 standard deviation, and the empirical probabilities of observations falling out of the range of the average ± 1.96 standard deviations.

In the following tables, "emp. prob. 1" is the empirical probability of observations falling out of the ± 1.96 standard deviations around the true value, whereas "emp. prob. 2" the empirical probability of observations falling out of the ± 1.96 standard deviations around the average.

Table 21 - Contrasts Between Different Cluster Sizes for Variance Components

	$D_{00} = 2$		$D_{01} = .2$		$D_{11} = .75$	
cluster size	200	2000	200	2000	200	2000
average	1.9187	1.9375	.1733	.1766	.7416	.7343
bias	-.0813	-.0625	-.0267	-.0234	-.0084	-.0157
variance	.0715	.0070	.0165	.0016	.0199	.0017
S. D.	.2674	.0839	.1285	.0404	.1412	.0413
emp. prob. 1	.08	.09	.07	.08	.055	.065
emp. prob. 2	.06	.06	.05	.045	.055	.06

Table 22 - Contrasts Between Different Cluster Sizes for Fixed Effects

	$\beta_{00} = -.5084$		$\beta_{01} = 1$		$\beta_{10} = 1$	
cluster size	200	2000	200	2000	200	2000
average	-.5107	-.5108	.9844	.9742	.9861	.9909
bias	-.0023	-.0024	-.0156	-.0258	-.0139	-.0091
variance	.0163	.0022	.0123	.0012	.0061	.0007
S. D.	.1278	.0473	.1107	.0342	.0778	.0255
emp. prob. 1	.045	.06	.075	.11	.06	.065
emp. prob. 2	.04	.065	.06	.045	.05	.05

Tables 21 and 22 showed that most of the estimates had fairly small amounts of negative biases, but the biases did not go down for data sets with a larger number of clusters. In terms of percentage, $D_{01} = .2$ had more than 10% negative biases for both

cluster sizes. The similar amounts of variation in D_{01} =.2 and D_{11} =.75 signified that Laplace6 had difficulty estimating D_{01}, a phenomenon the same with the other methods (Gauss and PQL2) in the previous model.

Histograms of the estimates not presented here showed that all of the estimates were fairly normally distributed. In effect, the empirical probability 2 (emp. prob. 2) also showed that percentages of observations falling out of the 95% interval around the average were around 5 for all estimates.

The variances of estimates using 200 clusters were all approximately 10 times larger than their counterparts using 2000 clusters. This indicated that as the number of clusters increases, the variation (variance) of Laplace6 estimates would decrease in proportion to the number of clusters. Thus, finally the estimates would peak at one point. Nevertheless, this peak would be a little off the true parameter, due to the negative bias. The empirical probability 1 also showed that a little more than 5% of the estimates were cast out of the 95% confidence interval around the true value. This was coherent with the finding of small negative biases. Since the estimates had a negative bias, the sample mean was shifted a little to the left of the true value, assuming both the sample mean and the true value were both normally distributed and had the same standard deviation. Thus, more estimates at the lower end than at the upper end of the empirical distribution would be rejected as plausible values from the distribution of the true value.

Chapter 6

DISCUSSION AND CONCLUSION

This dissertation uses Laplace's approximation method to solve the problems encountered in multilevel logistic models. In the process, I first deduced the multivariate Taylor expansion for use in expanding Laplace approximation to multivariate situations. Secondly, I derived the six moments of a multivariate normal distribution through its moment generating function. Then, I found the analogy between univariate moments and multivariate moments in doing Laplace's method. Using the above findings, I obtained the marginal likelihood of the multilevel logistic regression models as a simplified, scalar function of matrices. In finding maximum likelihood estimates of the fixed effects and the variance components of the random effects, I used implicit differential to take into consideration the dependence of the current estimate of the random effects on the parameters of interest. The result is the Laplace6 program in HLM (Bryk, Raudenbush, and Congdon, 1996), using as starting values the converged estimates by PQL.

Both univariate and bivariate random effects simulation studies and a real data analysis were carried out to evaluate Laplace6. The estimates were compared to those by the approximate maximum likelihood method using Gauss-Hermite Quadrature (Gauss) (MIXOR, 1993, 1994, 1997), the method of second-order Taylor expansion around the

conditional expectation (PQL2) (Goldstein and Rasbash, 1996), and PQL (Breslow and Clayton, 1993; Raudenbush, 1993).

The analysis of Thailand data is an example of how the multilevel Bernoulli model can be used to understand how student background, school background and national programs, such as pre-primary education, interact and affect educational outcome --- grade repetition. It was found that girls had a smaller risk of repeating a grade, that students with better nutrition (breakfast) had a lower risk, and that a higher family socioeconomic status and richer school resources (SES and enrollment) helped reduce the risk. However, there were interactions between student SES, student nutrition (breakfast) and school SES and school resources (log_enrollment), respectively. They indicated that students with poorer family background took more advantage of richer school resources than students with better family background in reducing the risk of repeating a grade. Moreover, all the background factors controlled, students having had pre-primary education still had a significantly lowered risk of repeating grades. The effect of the pre-primary education did not vary across different schools. Therefore, pre-primary education did have a positive effect in preparing children for primary schools.

Through using 4 different numbers of quadrature points in analyzing Thailand data, Laplace6 results were found to be more similar to those by larger numbers of points, i.e., 40 or 30 points, than all the other methods. Thus, Laplace6 seems to be a pretty accurate approximation to the marginal Bernoulli likelihood with a normal prior.

The extensive univariate simulation study indicated that all the programs except PQL performs reasonably well, with small biases at times, although PQL2 tend to have

slightly larger mean squared errors than Gauss-10 and Laplace6. Its efficiencies for the variance estimates especially trailed behind those of Laplace6. Laplace6 had the highest efficiency relative to all the methods for most of the parameters in all six univariate models, although the difference between the results of Laplace6 and those of Gauss-10 was small.

However, in the bivariate model with 100 data sets, the similarities in mean squared errors among Gauss-10 and Laplace6 disappeared. Laplace6 performed even better than in the univariate cases. While Gauss-10, Gauss20 and Laplace6 were all approximately unbiased, Laplace6 had the highest efficiencies over all the methods for most of the parameters. Its performance was even better than that of Gauss-20, which in turn was better than that of Gauss-10. PQL and PQL2 both had much smaller efficiencies than Laplace6.

The eighth model shows that Laplace6 estimates were normally distributed, and had a small amount of negative bias. However, the variance of the estimates did go down in proportion to the second level sample size. Thus, the approximate maximum likelihood estimates produced by Laplace6 indeed are approximately consistent estimates.

The analysis of Thailand data raises a question of how different the programs are and how much more useful Laplace6 or Gauss-Hermite Quadrature is for practice. The suggestion is, for univariate random effect models, PQL2 may do as well as Gauss-Hermite Quadrature and Laplace6; for multivariate random effects models, nevertheless, PQL2 may not give as good results as the latter two programs. On the other hand, PQL has serious negative bias for large variances of the random effects. Although in the

current model for Thailand data all programs happened to have the same amount of predictors that were significant at .05 level, it is still likely that for other models, PQL or PQL2 will have very different, and wrong, conclusions, originating from its negative bias in the parameters. Therefore, it is of both theoretical interest and of practical usefulness to have programs such as Gauss-Hermite Quadrature or Laplace6.

The advantage of Gaussian Quadrature lies in its flexibility in that the estimates can be found as accurately as the user wishes by just giving a larger number of quadrature points. Laplace approximation can go as accurate as one wishes, too, but it can be done only by the programmer, not the user.

The time needed for computation is a big advantage of Laplace6 over Gauss-Hermite Quadrature. As the experience with several random samples of the 100 bivariate data sets shows, Laplace6 was much faster than Gauss-Hermite Quadrature with 10 quadrature points specified, which was in turn much faster than Gauss-Hermite Quadrature with 20 points. However, given the exploration with different data sets here, 10 quadrature points could barely produce estimates as accurate as Laplace6. The time needed for Gauss-Hermite Quadrature to produce sufficiently accurate results will thus be much longer than for Laplace6.

Therefore, for educational research that is interested in dichotomous responses, such as grade repetition, high school dropout, or college admission, and that often is longitudinal and/or nested designs, Laplace6 is an accurate tool that is fast to converge. Although currently Laplace6 is available only for 2 level modeling, it should be straightforward to extend it to 3 level models.

On the other hand, the contribution of Laplace6 to the field of applied statistics/mathematics lies in the foundation upon which it is built, i.e., the multivariate Taylor series and the parallelism between moments in univariate normal distribution and those in multivariate normal distribution in applying the extended Laplace approximation. I expect that the method can be applied to give pretty accurate solutions to problems concerning integrals that have to be evaluated numerically.

APPENDICES

PRE-APPENDIX

PRE-APPENDIX

Formulae and Lemmas for Appendices A to D

- $dABC = (C^T \otimes A)dvecB$ (PA-1)

- Lemma 1 For a matrix F, suppose a is a scalar function of t, F does not involve t.

$$\frac{\partial\, vec(aF)}{\partial\, t^T} = vec(F) \otimes \frac{\partial\, a}{\partial\, t^T}. \tag{PA-2}$$

Proof Assume $F = \begin{bmatrix} f_{11}\ f_{12} \cdot \cdot \cdot \cdot \cdot\ f_{1n} \\ f_{21}\ f_{22} \cdot \cdot \cdot \cdot \cdot\ f_{2n} \\ \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \\ f_{m1}\ f_{m2} \cdot \cdot \cdot \cdot\ f_{mn} \end{bmatrix}$, and $aF = \begin{bmatrix} af_{11}\ af_{12} \cdot \cdot \cdot \cdot \cdot\ af_{1n} \\ af_{21}\ af_{22} \cdot \cdot \cdot \cdot \cdot\ af_{2n} \\ \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \\ af_{m1}\ af_{m2} \cdot \cdot \cdot \cdot\ af_{mn} \end{bmatrix}$

According to Magnus, we vectorize the matrix first before we take derivative of it with respect to a row vector. Therefore,

$$\frac{\partial aF}{\partial t^T} = \begin{bmatrix} \frac{\partial a}{\partial t^T} f_{11} \\ \frac{\partial a}{\partial t^T} f_{21} \\ . \\ . \\ \frac{\partial a}{\partial t^T} f_{m1} \\ \frac{\partial a}{\partial t^T} f_{12} \\ \frac{\partial a}{\partial t^T} f_{22} \\ . \\ . \\ \frac{\partial a}{\partial t^T} f_{m2} \\ . \\ . \\ . \\ \frac{\partial a}{\partial t^T} f_{mn} \end{bmatrix} = [vec(F) \otimes \frac{\partial a}{\partial t^T}].$$

Since F is not a function of t, each element in F becomes a scalar to the row vector $\frac{\partial a}{\partial t^T}$.

- For any two column vectors a, b, $a \otimes b^T = ab^T$ (PA-3)

For $F(X,Y) = X \otimes Y$, X is an $n \times q$ matrix, and Y a $p \times r$ matrix.

Then $dF(X,Y) = vec(dX \otimes Y) + vec(X \otimes dY)$,

where $vec(dX \otimes Y) = (I_q \otimes K_{rn} \otimes I_p)(I_{nq} \otimes vecY)dvecX$ (PA-4)

$$vec(X \otimes dY) = (I_q \otimes K_{rn} \otimes I_p)(vecX \otimes I_{pr})dvecY,$$ (PA-5)

I_n being an $n \times n$ identity matrix, and

K_n being an $n \times n$ commutation matrix. (Magnus and Neudecker(1988), p. 188)

- Suppose a is a $p \times 1$ vector and b, a $q \times 1$ vector. Then $a \otimes b^T$ is a $p \times q$ matrix and so is $b^T \otimes a$. Thus $a \otimes b^T = b^T \otimes a$. (PA-6)
- Suppose $s = a^T b$, a and b are $n \times 1$ vectors. Then $s = tr(b^T a) = tr(ba^T)$. (PA-7)
- For same-order matrices, A and B, $tr(A^T B) = (vecA)^T (vecB)$. (PA-8)
- Two vectors a and b, $ab^T = (a \otimes b^T)$ (PA-9)
- For two vectors a and b, $vec(ab^T) = (b \otimes a)$. (PA-10)
- $A \otimes B \otimes C = (A \otimes B) \otimes C = A \otimes (B \otimes C)$ (PA-11)
- Lemma 2 For a p × q matrix A, an s × 1 column vector c,

$$vec[A \otimes c] = vec(A) \otimes c. \quad \text{(PA-12)}$$

Proof

$$vec(A \otimes c) = vec\begin{bmatrix} a_{11}\begin{bmatrix} c_1 \\ \vdots \\ c_s \end{bmatrix} & a_{12}\begin{bmatrix} c_1 \\ \vdots \\ c_s \end{bmatrix} & \cdots & a_{1q}\begin{bmatrix} c_1 \\ \vdots \\ c_s \end{bmatrix} \\ & \vdots & & \\ a_{p1}\begin{bmatrix} c_1 \\ \vdots \\ c_s \end{bmatrix} & a_{p2}\begin{bmatrix} c_1 \\ \vdots \\ c_s \end{bmatrix} & \cdots & a_{pq}\begin{bmatrix} c_1 \\ \vdots \\ c_s \end{bmatrix} \end{bmatrix} = vec(A) \otimes c .$$

- If A and B are square matrices, $tr(A \otimes B) = tr(A)tr(B)$. (PA-13)
- If AB and CD exist, then $(A \otimes B)(C \otimes D) = AC \otimes BD$. (PA-14)
- If AB is a square matrix, then $tr(AB) = tr(BA)$. (PA-15)

- $vec(ABC) = (C^T \otimes A)vecB$ (PA-16)

APPENDIX A

APPENDIX A

Multivariate Taylor Series Expansion

This appendix will derive the following theorem:

Theorem: The *m*-th order approximation in n-variable Taylor series is $\frac{1}{m!}(\overset{m-1}{\underset{1}{\otimes}} \vec{b}^T) f^{(m)} \vec{b}$, where $\vec{b}$ is an $n \times 1$ vector; $f^{(m)}$ is the *m*-th derivative of the function *f*, *f* a function of $\vec{b}$; the derivative is obtained by first vectorizing the *(m-1)-th* derivative and then differentiating with respect to $\vec{b}^T$; $\overset{m-1}{\underset{1}{\otimes}} \vec{b}^T$ is the Kronecker product of $\vec{b}^T$, repeated *m-1* times.

Proof

Fulks (Fulks, 1978, p.331) has the following multivariate Taylor series theorem.

Theorem. Let *f* and all its partial derivatives up through order *n* be continuous in a neighborhood *N* of Q_0. Then for *P* in *N*,

$$f(P) = f(Q_0) + [(P-Q_0)\bullet\nabla]f(Q)\big|_{Q=Q_0} + \frac{1}{2!}[(P-Q_0)\bullet\nabla]^2 f(Q)\big|_{Q=Q_0} + \ldots + \frac{1}{(n-1)!}[(P-Q_0)\bullet\nabla]^{n-1} f(Q)\big|_{Q=Q_0} + \frac{1}{n!}[(P-Q_0)\bullet\nabla]^n f(Q)\big|_{Q=P_0} \quad (1)$$

(where $A \bullet B$ denotes the inner product of *A* and *B*; P_0 is a point on the segment connecting P to Q_0; the symbol ∇ indicates the differentiation operation.)

Given (1) then, for $Q_0 = (a_1, a_2, \ldots, a_n)$ and $P = (x_1, x_2, \ldots, x_n)$, Taylor expansion of $f(x_1, x_2, \ldots, x_n)$ up to the fourth order is:

$$f(x_1,x_2,\ldots,x_n)=f(a_1,a_2,\ldots,a_n)+[(x_1-a_1)\frac{\partial}{\partial x_1}+(x_2-a_2)\frac{\partial}{\partial x_2}+\ldots$$

$$+(x_n-a_n)\frac{\partial}{\partial x_n}]f(a_1,a_2,\ldots,a_n)$$

$$+\frac{1}{2!}[(x_1-a_1)\frac{\partial}{\partial x_1}+(x_2-a_2)\frac{\partial}{\partial x_2}+\ldots+(x_n-a_n)\frac{\partial}{\partial x_n}]^2 f(a_1,a_2,\ldots,a_n)$$

$$+\frac{1}{3!}[(x_1-a_1)\frac{\partial}{\partial x_1}+(x_2-a_2)\frac{\partial}{\partial x_2}+\ldots+(x_n-a_n)\frac{\partial}{\partial x_n}]^3 f(a_1,a_2,\ldots,a_n) \qquad (2)$$

$$+\frac{1}{4!}[(x_1-a_1)\frac{\partial}{\partial x_1}+(x_2-a_2)\frac{\partial}{\partial x_2}+\ldots+(x_n-a_n)\frac{\partial}{\partial x_n}]^4 f(a_1,a_2,\ldots,a_n)+\ \ldots$$

$$+\frac{1}{n!}[(x_1-a_1)\frac{\partial}{\partial x_1}+(x_2-a_2)\frac{\partial}{\partial x_2}+\ldots+(x_n-a_n)\frac{\partial}{\partial x_n}]^n f(a_1,a_2,\ldots,a_n).$$

We look more closely starting at the second order ignoring the factorial:

$$[(x_1-a_1)\frac{\partial}{\partial x_1}+(x_2-a_2)\frac{\partial}{\partial x_2}+\ldots+(x_n-a_n)\frac{\partial}{\partial x_n}]^2 f(a_1,a_2,\ldots,a_n)$$

$$=\Bigg[\sum_{i=1}^{n}(x_1-a_1)(x_i-a_i)\frac{\partial^2 f(a_1,a_2,\ldots,a_n)}{\partial x_1\partial x_i}$$

$$+\sum_{i=1}^{n}(x_2-a_2)(x_i-a_i)\frac{\partial^2 f(a_1,a_2,\ldots,a_n)}{\partial x_2\partial x_i}+\ldots$$

$$+\sum_{i=1}^{n}(x_n-a_n)(x_i-a_i)\frac{\partial^2 f(a_1,a_2,\ldots,a_n)}{\partial x_n\partial x_i}\Bigg]$$

$$=\sum_{i,j}^{n}(x_i-a_i)(x_j-a_j)\frac{\partial^2 f(a_1,a_2,\ldots,a_n)}{\partial x_i\partial x_j}; \qquad (3)$$

the third order, ignoring the factorial:

$$[(x_1-a_1)\frac{\partial}{\partial x_1}+(x_2-a_2)\frac{\partial}{\partial x_2}+\ldots+(x_n-a_n)\frac{\partial}{\partial x_n}]^3 f(a_1,a_2,\ldots,a_n)$$

$$=[(x_1-a_1)\frac{\partial}{\partial x_1}+(x_2-a_2)\frac{\partial}{\partial x_2}+\ldots+(x_n-a_n)\frac{\partial}{\partial x_n}]^2\times$$

$$[(x_1-a_1)\frac{\partial}{\partial x_1}+(x_2-a_2)\frac{\partial}{\partial x_2}+\ldots+(x_n-a_n)\frac{\partial}{\partial x_n}]f(a_1,a_2,\ldots,a_n)$$

$$= \left[\sum_{i,j}^{n} (x_1 - a_1)(x_i - a_i)(x_j - a_j) \frac{\partial^3 f(a_1, a_2, \ldots, a_n)}{\partial x_i \partial x_j \partial x_1} \right.$$

$$+ \sum_{i,j}^{n} (x_2 - a_2)(x_i - a_i)(x_j - a_j) \frac{\partial^3 f(a_1, a_2, \ldots, a_n)}{\partial x_i \partial x_j \partial x_2} + \ldots$$

$$\left. + \sum_{i,j}^{n} (x_n - a_n)(x_i - a_i)(x_j - a_j) \frac{\partial^3 f(a_1, a_2, \ldots, a_n)}{\partial x_i \partial x_j \partial x_n} \right]$$

$$= \sum_{i,j,k}^{n} (x_k - a_k)(x_i - a_i)(x_j - a_j) \frac{\partial^3 f(a_1, a_2, \ldots, a_n)}{\partial x_i \partial x_j \partial x_k}; \qquad (4)$$

and the fourth order, also ignoring the factorial:

$$\left[(x_1 - a_1) \frac{\partial}{\partial x_1} + (x_2 - a_2) \frac{\partial}{\partial x_2} + \ldots + (x_n - a_n) \frac{\partial}{\partial x_n} \right]^4 f(a_1, a_2, \ldots, a_n)$$

$$= \left[(x_1 - a_1) \frac{\partial}{\partial x_1} + (x_2 - a_2) \frac{\partial}{\partial x_2} + \ldots + (x_n - a_n) \frac{\partial}{\partial x_n} \right]^3 \times$$

$$\left[(x_1 - a_1) \frac{\partial}{\partial x_1} + (x_2 - a_2) \frac{\partial}{\partial x_2} + \ldots + (x_n - a_n) \frac{\partial}{\partial x_n} \right] f(a_1, a_2, \ldots, a_n)$$

$$= \left[\sum_{i,j,k}^{n} (x_1 - a_1)(x_i - a_i)(x_j - a_j)(x_k - a_k) \frac{\partial^4 f(a_1, a_2, \ldots, a_n)}{\partial x_i \partial x_j \partial x_k \partial x_1} \right.$$

$$+ \sum_{i,j,k}^{n} (x_2 - a_2)(x_i - a_i)(x_j - a_j)(x_k - a_k) \frac{\partial^4 f(a_1, a_2, \ldots, a_n)}{\partial x_i \partial x_j \partial x_k \partial x_2} + \ldots$$

$$\left. + \sum_{i,j,k}^{n} (x_n - a_n)(x_i - a_i)(x_j - a_j)(x_k - a_k) \frac{\partial^4 f(a_1, a_2, \ldots, a_n)}{\partial x_i \partial x_j \partial x_k \partial x_n} \right]$$

$$= \sum_{i,j,k,l}^{n} (x_l - a_l)(x_i - a_i)(x_j - a_j)(x_k - a_k) \frac{\partial^4 f(a_1, a_2, \ldots, a_n)}{\partial x_i \partial x_j \partial x_k \partial x_l}. \qquad (5)$$

I will arrange each order of the above into scalar functions of matrices. In taking derivatives, I will always follow what Magnus preaches, that "the only sensible definition of a matrix derivative" is to vectorize the matrix first and then take its derivative with respect to a row vector. Thus whether the original function is a row

vector, a column vector or a matrix, to differentiate it we will always vectorize it first, using the formula: $vec(ABC) = (C^T \otimes A)vecB$, and then take the derivative.

To simplify notation, let

$$b = \begin{bmatrix} x_1 - a_1 \\ x_2 - a_2 \\ . \\ . \\ x_n - a_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ b_n \end{bmatrix}, \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ x_n \end{bmatrix} = z \text{, and } \frac{\partial f(a_1, a_2, \ldots, a_n)}{\partial x_i} = f_{x_i},$$

$\frac{\partial^2 f(a_1, a_2, \ldots, a_n)}{\partial x_i \partial x_j} = f_{x_i x_j}$, etc.. Thus, for example, the first derivative is

$$f^{(1)}(x_1, x_2, \ldots, x_n) = \frac{\partial f}{\partial z^T} = \begin{bmatrix} f_{x_1} & f_{x_2} & \cdots & f_{x_n} \end{bmatrix}.$$

Then the first order can be rewritten as

$$f^{(1)}b = \begin{bmatrix} f_{x_1} & f_{x_2} & .. & f_{x_n} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ b_n \end{bmatrix} = f_{x_1} b_1 + f_{x_2} b_2 + .. + f_{x_n} b_n$$

$$= (x_1 - a_1) \frac{\partial f(a_1, a_2, \ldots, a_n)}{\partial x_1} + (x_2 - a_2) \frac{\partial f(a_1, a_2, \ldots, a_n)}{\partial x_2}$$

$$+ \ldots + (x_n - a_n) \frac{\partial f(a_1, a_2, \ldots, a_n)}{\partial x_n}.$$

The second derivative of f is obtained by taking derivative of vec $f^{(1)}$ with respect to z^T.

$$f^{(2)} = \frac{\partial\, vecf^{(1)}}{\partial\, z^T} = \begin{bmatrix} & & \\ & & \\ f_{x_1x_1} & f_{x_1x_2} & \cdots & f_{x_1x_n} \\ f_{x_2x_1} & f_{x_2x_2} & \cdots & f_{x_2x_n} \\ & \cdot & & \\ & \cdot & & \\ & \cdot & & \\ f_{x_nx_1} & f_{x_nx_2} & \cdots & f_{x_nx_n} \end{bmatrix}.$$

Then the second order term without the factorial is derived by differentiating the first order term, $f^{(1)}b$, and then post-multiplying it with b: since $f^{(1)}$ is $1 \times n$, in applying Equation PA-1, we regard $f^{(1)}$ as the "B", then "A" is a 1×1 scalar of 1, and b as the "C". So that we have

$\frac{\partial}{\partial z^T} f^{(1)}b = b^T f^{(2)}$. Therefore,

$$b^T f^{(2)} b = \begin{bmatrix} b_1 & b_2 & \cdots & b_n \end{bmatrix} \begin{bmatrix} & & \\ & & \\ f_{x_1x_1} & f_{x_1x_2} & \cdots & f_{x_1x_n} \\ f_{x_2x_1} & f_{x_2x_2} & \cdots & f_{x_2x_n} \\ & \cdot & & \\ & \cdot & & \\ & \cdot & & \\ f_{x_nx_1} & f_{x_nx_2} & \cdots & f_{x_nx_n} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ b_n \end{bmatrix}$$

$$= \left[\sum_{i=1}^{n} b_i f_{x_i x_1} \;\; \sum_{i=1}^{n} b_i f_{x_i x_2} \;\; \cdots \;\; \sum_{i=1}^{n} b_i f_{x_i x_n} \right] \begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ b_n \end{bmatrix} = \sum_{i=1}^{n} b_1 b_i f_{x_i x_1} + \sum_{i=1}^{n} b_2 b_i f_{x_i x_2} + \ldots + \sum_{i=1}^{n} b_n b_i f_{x_i x_n}$$

$$= \sum_{i,j}^{n} b_j b_i f_{x_i x_j} = \sum_{i,j}^{n} (x_i - a_i)(x_j - a_j) \frac{\partial^2 f(a_1, a_2, \ldots, a_n)}{\partial x_i \partial x_j}$$, which is Equation 3.

Similarly, the third derivative is obtained by vectorizing $f^{(2)}$, and then taking its derivative with respect to a row vector z^T: (the dashed lines are to make it easier for the reader to see the original vectors from $f^{(2)}$)

$$f^{(3)} = \frac{\partial}{\partial z^T} vecf^{(2)} = \frac{\partial}{\partial z^T} \begin{bmatrix} f_{x_1x_1} \\ f_{x_2x_1} \\ . \\ . \\ f_{x_nx_1} \\ --- \\ f_{x_1x_2} \\ f_{x_2x_2} \\ . \\ . \\ f_{x_nx_2} \\ --- \\ . \\ . \\ . \\ --- \\ f_{x_1x_n} \\ f_{x_2x_n} \\ . \\ . \\ f_{x_nx_n} \end{bmatrix} = \begin{bmatrix} f_{x_1x_1x_1} & f_{x_1x_1x_2} & .. & f_{x_1x_1x_n} \\ f_{x_2x_1x_1} & f_{x_2x_1x_2} & .. & f_{x_2x_1x_n} \\ . & & & \\ . & & & \\ f_{x_nx_1x_1} & f_{x_nx_1x_2} & .. & f_{x_nx_1x_n} \\ - & - & - & - \\ f_{x_1x_2x_1} & f_{x_1x_2x_2} & .. & f_{x_1x_2x_n} \\ f_{x_2x_2x_1} & f_{x_2x_2x_2} & .. & f_{x_2x_2x_n} \\ . & & & \\ . & & & \\ f_{x_nx_2x_1} & f_{x_nx_2x_2} & .. & f_{x_nx_2x_n} \\ - & - & - & - \\ . & & & \\ . & & & \\ . & & & \\ - & - & - & - \\ f_{x_1x_nx_1} & f_{x_1x_nx_2} & .. & f_{x_1x_nx_n} \\ f_{x_2x_nx_1} & f_{x_2x_nx_2} & .. & f_{x_2x_nx_n} \\ . & & & \\ . & & & \\ f_{x_nx_nx_1} & f_{x_nx_nx_2} & .. & f_{x_nx_nx_n} \end{bmatrix}$$

Then the third order term without the factorial can be obtained by differentiating the second order term, $b^T f^{(2)} b$, and then post-multiplying it with b.

By Equation PA-1 again, $(db^T f^{(2)} b)b = (b^T \otimes b^T) f^{(3)} b$

$$= \begin{bmatrix} b_1b_1 & b_1b_2 & .. & b_1b_n & b_2b_1 & b_2b_2 & .. & b_2b_n & .... & b_nb_1 & b_nb_2 & .. & b_nb_n \end{bmatrix} \begin{bmatrix} f_{x_1x_1x_1} & f_{x_1x_1x_2} & .. & f_{x_1x_1x_n} \\ f_{x_2x_1x_1} & f_{x_2x_1x_2} & .. & f_{x_2x_1x_n} \\ . & & & \\ . & & & \\ f_{x_nx_1x_1} & f_{x_nx_1x_2} & .. & f_{x_nx_1x_n} \\ \hline f_{x_1x_2x_1} & f_{x_1x_2x_2} & .. & f_{x_1x_2x_n} \\ f_{x_2x_2x_1} & f_{x_2x_2x_2} & .. & f_{x_2x_2x_n} \\ . & & & \\ . & & & \\ f_{x_nx_2x_1} & f_{x_nx_2x_2} & .. & f_{x_nx_2x_n} \\ \hline . & & & \\ . & & & \\ . & & & \\ \hline f_{x_1x_nx_1} & f_{x_1x_nx_2} & .. & f_{x_1x_nx_n} \\ f_{x_2x_nx_1} & f_{x_2x_nx_2} & .. & f_{x_2x_nx_n} \\ . & & & \\ . & & & \\ f_{x_nx_nx_1} & f_{x_nx_nx_2} & .. & f_{x_nx_nx_n} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ b_n \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i,j}^{n} b_ib_jf_{x_ix_jx_1} & \sum_{i,j}^{n} b_ib_jf_{x_ix_jx_2} & .. & \sum_{i,j}^{n} b_ib_jf_{x_ix_jx_n} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ b_n \end{bmatrix}$$

$$= \sum_{i,j}^{n} b_1b_ib_jf_{x_ix_jx_1} + \sum_{i,j}^{n} b_2b_ib_jf_{x_ix_jx_2} + .. + \sum_{i,j}^{n} b_nb_ib_jf_{x_ix_jx_n} = \sum_{i,j,k}^{n} b_kb_ib_jf_{x_ix_jx_k}$$

$$= \sum_{i,j,k}^{n} (x_k - a_k)(x_i - a_i)(x_j - a_j) \frac{\partial^3 f(a_1, a_2, \ldots, a_n)}{\partial x_i \partial x_j \partial x_k}$$, which is Equation 4.

The fourth order term can be expressed in terms of a matrix and vectors in the way exactly as what we did for the second- and third- order terms. The lay-out is omitted here because of its tedium, but it can be expressed as, with $f^{(4)} = \frac{\partial}{\partial z^T} f^{(3)}$,

$$[d(b^T \otimes b^T) f^{(3)} b] b = (b^T \otimes b^T \otimes b^T) f^{(4)} b$$

$$= \sum_{i,j,k,l}^{n} (x_l - a_l)(x_i - a_i)(x_j - a_j)(x_k - a_k) \frac{\partial^4 f(a_1, a_2, \ldots, a_n)}{\partial x_i \partial x_j \partial x_k \partial x_l}$$, which is equal to Equation 5.

Therefore, I have proved the theorem.

APPENDIX B

APPENDIX B

The Six Moments of a Multivariate Normal Distribution

Introduction

This Appendix finds the first six moments of a multivariate normal distribution. Since methods are routinely applied, I will prove up to the fourth moment. The last two moments will be proved with minimal elaboration.

The moment generating function of a multivariate normal distribution with mean 0 and variance Σ, N(0, Σ), is $\exp\left(\frac{1}{2}t^T\Sigma\ t\right)$. Suppose the dimension is q. Then t is a $q \times 1$ vector, Σ is a $q \times q$ matrix, and $\exp\left(\frac{1}{2}t^T\Sigma\ t\right)$ is a scalar. The exponential will remain no matter how many times the derivative is taken with respect to the elements inside it. For simplicity, define $m = \exp\left(\frac{1}{2}t^T\Sigma\ t\right) = f(t)$.

First derivative

Theorem 1 $$f^{(1)}(t) = \frac{\partial}{\partial t^T} m = m\frac{\partial}{\partial t^T}\left(\frac{1}{2}t^T\Sigma\ t\right) = mt^T\Sigma \qquad (1)$$

Proof

Because of the exponential function, $\frac{\partial}{\partial t^T} m = m\frac{\partial}{\partial t^T}\left(\frac{1}{2}t^T\Sigma\ t\right)$.

For $\frac{\partial}{\partial t^T}\left(\frac{1}{2}t^T\Sigma\, t\right)$, I first take derivative with respect to t, the last term, and then t^T, the first term. To take derivative with respect to t, I regard $t^T\Sigma$ as A, and t as B in Equation PA-1. For C in Equation PA-1, I use an identity matrix of the same dimension as the column of t, which is 1. Therefore, the identity matrix in this case is a scalar 1. So I have

$$(1\otimes\frac{1}{2}t^T\Sigma)\frac{\partial t}{\partial t^T}=\frac{1}{2}t^T\Sigma. \tag{2}$$

To take derivative with respect to t^T, I regard t^T as B, and $\Sigma\, t$ as C also in (PA-1). For A I use an identity matrix of the same dimension as the row of $t^T\Sigma$, which is also 1. Therefore, the identity matrix in this case is a scalar, 1. So I have

$$[\frac{1}{2}(\Sigma\, t)^T\otimes 1]\frac{\partial\, vec(t^T)}{\partial t^T}=\frac{1}{2}t^T\Sigma. \tag{3}$$

Adding up (2) and (3), $\frac{\partial}{\partial t^T}m=m\frac{\partial}{\partial t^T}\left(\frac{1}{2}t^T\Sigma\, t\right)=\frac{m}{2}(2t^T\Sigma)=mt^T\Sigma$ (4)

The first moment is 0, obtained by setting $t = 0$.

Second derivative

To find the second derivative, I transpose (4) and then take derivative with respect to the row vector t^T again. That is, $\frac{\partial^2}{\partial t^T\partial t}m=\frac{\partial}{\partial t^T}m\Sigma\, t$.

Theorem 2 $\quad f^{(2)}(t)=\frac{\partial^2}{\partial t^T\partial t}m=\frac{\partial}{\partial t^T}m\Sigma\, t=m\Sigma+m\Sigma\, t\, t^T\Sigma$ (5)

Proof I take derivative first with respect to m using Equation PA-2, and then with respect to t, and finally I sum up the two parts.

To take derivative with respect to m, regard $\Sigma\ t$ as F in Equation PA-2. Then I have

$$m[vec(\Sigma\ t)\otimes t^T\Sigma] = m[(\Sigma\ t)\otimes t^T\Sigma] = m\Sigma\ t\ t^T\Sigma, \qquad (6)$$

by Equation PA-3, taking advantage of the fact that both $\Sigma\ t$ and $t^T\Sigma$ are vectors.

To take derivative with respect to t, regard $m\Sigma$ as A, and t as B in Equation PA-1. Then again I have 1 as C, since the column dimension of t is 1.

The derivative is $[1\otimes(m\Sigma)]\dfrac{\partial t}{\partial t^T} = m\Sigma$. (7)

So Equation 5 is proved by adding up (6) and (7).

Again, setting $t = 0$, I have the second moment, Σ.

Third derivative

Theorem 3 $$f^{(3)}(t) = \frac{\partial\ vec\dfrac{\partial^2 m}{\partial t^T \partial t}}{\partial t^T} \qquad (8)$$
$$= mvec(\Sigma)t^T\Sigma + mvec(\Sigma\ t\ t^T\Sigma)t^T\Sigma + m[\Sigma\otimes(\Sigma t)] + m[(\Sigma t)\otimes\Sigma]$$

Proof

$$\text{With } f^{(3)}(t) = \frac{\partial}{\partial t^T} vecf^{(2)}(t) = \frac{\partial}{\partial t^T} vec(m\Sigma + m\Sigma\ t\ t^T\Sigma), \qquad (9)$$

I will differentiate the first item at the right hand side first and then the second term.

For the first term, according to Equation PA-2, since only m is a function of t, the derivative is

$$\frac{\partial}{\partial t^T} vec(m\Sigma) = [vec(\Sigma)\otimes mt^T\Sigma] = m[vec(\Sigma)\otimes t^T\Sigma] = \text{mvec}(\Sigma)\text{t}^\text{T}\Sigma, \qquad (10)$$

by Equation PA-3.

For $\frac{\partial}{\partial t^T} vec(m\ \Sigma\ t\ t^T\Sigma)$, I first take derivative with respect to m, then t and finally t^T.

With respect to m, I have $[vec(\Sigma\ t\ t^T\Sigma)\otimes mt^T\Sigma] = mvec(\Sigma\ t\ t^T\Sigma)t^T\Sigma$. (11)

With respect to t, I have $t^T\Sigma$ as C in Equation PA-1, $m\Sigma$ as A, and t as B. So the derivative is $[(t^T\Sigma)^T\otimes(m\Sigma)]\frac{\partial t}{\partial t^T} = m[(\Sigma\ t)\otimes\Sigma]$, (12)

since Σ is a symmetric matrix.

With respect to t^T, regard Σ as C in Equation PA-1, $m\Sigma t$ as A, and t^T as B. Then the derivative is $[\Sigma\otimes(m\Sigma t)]\frac{\partial\ vec(t^T)}{\partial t^T} = m[\Sigma\otimes(\Sigma t)]\frac{\partial t}{\partial t^T} = m[\Sigma\otimes(\Sigma t)]$. (13)

The theorem is proved by summing up (10), (11), (12), and (13).

The third moment is 0 by setting t to 0.

Fourth derivative

Theorem 4

$$f^{(4)}(t) = \frac{\partial\ vec\dfrac{\partial^2 m}{\partial t^T\partial t}}{\partial t^T} = m(\Sigma\otimes vec(\Sigma)) + mvec[vec(\Sigma)t^T\Sigma]t^T\Sigma$$
$$+ m[vec(\Sigma\otimes\Sigma t)\otimes t^T\Sigma] + m[vec(\Sigma)\otimes\Sigma]$$
$$+ m[vec(\Sigma t\otimes\Sigma)\otimes t^T\Sigma] + m\{(K_{q,q}\otimes I_q)(\Sigma\otimes vec(\Sigma))\}$$
$$+ m\{vec[vec(\Sigma\ t\ t^T\Sigma)t^T\Sigma]\otimes t^T\Sigma\}$$
$$+ m\{(\Sigma\ t)\otimes I_{q^2}\}\ \{[(\Sigma\ t)\otimes\Sigma] + [\Sigma\otimes(\Sigma\ t)]\}$$
$$+ m[\Sigma\otimes vec(\Sigma\ t\ t^T\Sigma)] \quad (14)$$

Proof

$$f^{(4)}(t) = \frac{\partial}{\partial t^T} vec\{mvec(\Sigma)t^T\Sigma + m(\Sigma\otimes\Sigma t) + m(\Sigma t\otimes\Sigma)) + mvec(\Sigma\ t\ t^T\Sigma)t^T\Sigma\} \quad (15)$$

For the first item in the right hand side of the above equation, we have:

$$\frac{\partial}{\partial t^T} vec(mvec(\Sigma)t^T\Sigma) = m(\Sigma \otimes vec(\Sigma)) + m\{vec[vec(\Sigma)t^T\Sigma] \otimes t^T\Sigma\}$$

$$= m(\Sigma \otimes vec(\Sigma)) + mvec[vec(\Sigma)t^T\Sigma]t^T\Sigma \qquad (16)$$

The first item in the above equation results by taking derivative with respect to t^T. That is, Σ is regarded as C in Equation PA-1, $mvec(\Sigma)$ as A, and t^T as B. The second item results by taking derivative with respect to m by applying Equation PA-3.

For the second item in (15), the derivative is

$$\frac{\partial}{\partial t^T} vec(m(\Sigma \otimes \Sigma t)) = m[vec(\Sigma \otimes \Sigma t) \otimes t^T\Sigma] + m[vec(\Sigma) \otimes \Sigma] \qquad (17)$$

The first item in (17) results by taking derivative with respect to m, the second by taking derivative with respect to t. This second term is obtained by using Equation PA-5 for taking derivative of a matrix at the right hand side of a Kronecker product.

According to (PA-5), I have

$$\begin{aligned} &m\{(I_q \otimes K_{1q} \otimes I_q)[vec(\Sigma) \otimes I_q]\frac{\partial}{\partial t^T}[vec(\Sigma\, t)] \\ &= m\{(I_q \otimes K_{1q} \otimes I_q)[vec(\Sigma) \otimes I_q]\Sigma = m[vec(\Sigma) \otimes \Sigma], \end{aligned} \qquad (18)$$

where K_{1q} is a $q \times q$ commutation matrix, and thus also an identity matrix, I_q.

Therefore, $(I_q \otimes K_{1q} \otimes I_q) = (I_q \otimes I_q \otimes I_q) = I_{q^3}$ can be ignored. Furthermore,

$m[vec(\Sigma) \otimes I_q]\Sigma = m[vec(\Sigma) \otimes I_q](1 \otimes \Sigma) = m[vec(\Sigma) \otimes \Sigma]$, which is the second term in (17).

For the third item in (15), the derivative is

$$\frac{\partial}{\partial t^T} vec(m(\Sigma t \otimes \Sigma)) = m[vec(\Sigma t \otimes \Sigma) \otimes t^T \Sigma] + m(K_{qq} \otimes I_q)(\Sigma \otimes vec(\Sigma)) \qquad (19)$$

where K_{qq} is a $q^2 \times q^2$ commutation matrix. For the kth row in the K_{qq}, $(n-1)q < k \le nq$, $n = 1, ..., q$, only the $(n + \{k - [(n-1)q + 1]\}q)$-th term is 1, the others are 0.

The first item in the right hand side of (19) results by taking derivative with respect to m. The second term is derived using Equation PA-4 for taking derivative of a matrix at the left hand side of a Kronecker product. According to (PA-4), the derivative is

$$\begin{aligned} &m\{(1 \otimes K_{qq} \otimes I_q)(I_q \otimes vec(\Sigma))\} \frac{\partial}{\partial t^T} vec(\Sigma\, t) \\ &= m(K_{qq} \otimes I_q)(I_q \otimes vec(\Sigma))\Sigma = m(K_{qq} \otimes I_q)(\Sigma \otimes vec(\Sigma)), \end{aligned} \qquad (20)$$

for $(I_q \otimes vec(\Sigma))\Sigma = (I_q \otimes vec(\Sigma))(\Sigma \otimes 1) = (\Sigma \otimes vec(\Sigma))$. Equation 20 is the second term in Equation 19.

For the last term in (15), $mvec(\Sigma\ t\ t^T\Sigma)t^T\Sigma$, I take derivative with respect to m first, applying Equation PA-2; then t and t^T inside the *vec* function, and finally t^T outside the *vec* function. The derivative with respect to m is $m\{vec[vec(\Sigma\ t\ t^T\Sigma)t^T\Sigma] \otimes t^T\Sigma\}$. (21)

To find the derivative with respect to the t inside the *vec*, I use Equation PA-1 and the chain rule. First I regard $vec(\Sigma\ t\ t^T\Sigma)$ as B in (PA-1) and $t^T\Sigma$ as C, and then the A in (PA-1) is an identity matrix of the same dimension of the row vector as $mvec(\Sigma\ t\ t^T\Sigma)$, which is q^2. That is, I have $m\{(t^T\Sigma)^T \otimes I_{q^2}\} \dfrac{\partial\ vec(\Sigma\ t\ t^T\Sigma)}{\partial\ t^T}$. (22)

For $\frac{\partial\, vec(\Sigma\ t\ t^T\Sigma)}{\partial\, t^T}$, I apply (PA-1) again, which brings

$$\frac{\partial\, vec(\Sigma\ t\ t^T\Sigma)}{\partial\, t^T} = [(t^T\Sigma)^T \otimes \Sigma]\frac{\partial\, t}{\partial\, t^T} + [\Sigma \otimes (\Sigma\ t)]\frac{\partial\, vec(t^T)}{\partial\, t^T} \qquad (23)$$
$$= [(\Sigma\ t) \otimes \Sigma] + [\Sigma \otimes (\Sigma\ t)]$$

So the derivative with respect to t and t^T inside the *vec* function is, substituting (23) into

(22), $\quad m\{(t^T\Sigma)^T \otimes I_{q^2}\}\ \{[(\Sigma\ t) \otimes \Sigma] + [\Sigma \otimes (\Sigma\ t)]\} \qquad (24)$

Finally, to find the derivative with respect to t^T outside the *vec* function, I apply (PA-1), with $vec(\Sigma\ t\ t^T\Sigma)$ as A, t^T as B, and Σ as C. Then I have

$$m[\Sigma \otimes vec(\Sigma\ t\ t^T\Sigma)]\frac{\partial}{\partial\, t^T} vec(t^T) = m[\Sigma \otimes vec(\Sigma\ t\ t^T\Sigma)]. \qquad (25)$$

To sum up, the derivative of the last term in (15) is, adding (21),(24), and (25),

$$\begin{aligned}\frac{\partial}{\partial\, t^T}[mvec(\Sigma\ t\ t^T\Sigma)t^T\Sigma] = m\{vec[vec(\Sigma\ t\ t^T\Sigma)t^T\Sigma] \otimes t^T\Sigma\} \\ + m\{(t^T\Sigma)^T \otimes I_{q^2}\}\ \{[(\Sigma\ t) \otimes \Sigma] + [\Sigma \otimes (\Sigma\ t)]\} \\ + m[\Sigma \otimes vec(\Sigma\ t\ t^T\Sigma)]. \qquad (26)\end{aligned}$$

The theorem is proved by summing up (16), (17), (19) and (26).

The fourth moment results from setting t to 0 in all the items in the fourth derivative, Equation 14:

$$(\Sigma \otimes vec(\Sigma)) + [vec(\Sigma) \otimes \Sigma] + \{(K_{qq} \otimes I_q)(\Sigma \otimes vec(\Sigma))\}. \qquad (27)$$

Fifth Derivative

Theorem 5 The fifth derivative is

$$\frac{\partial vec\dfrac{\partial vec\dfrac{\partial vec\dfrac{\partial^2 m}{\partial t^T \partial t}}{\partial t^T}}{\partial t^T}}{\partial t^T} = m[vec(\Sigma \otimes vec(\Sigma)) \otimes t^T\Sigma]$$
$$+ mvec\{vec[vec(\Sigma)t^T\Sigma]t^T\Sigma\} \otimes (t^T\Sigma) + m(\Sigma \otimes vec(vec(\Sigma)t^T\Sigma)) + m(\Sigma t \otimes I_{q^3})(\Sigma \otimes vec(\Sigma))$$
$$+ m\{vec[vec(\Sigma \otimes \Sigma t) \otimes t^T\Sigma] \otimes t^T\Sigma\} + m(\Sigma \otimes vec(\Sigma \otimes \Sigma t)) + m(\Sigma t \otimes I_{q^3})(vec\Sigma \otimes \Sigma)$$
$$+ m(vec[vec(\Sigma) \otimes \Sigma] \otimes t^T\Sigma)$$
$$+ mvec[vec(\Sigma t \otimes \Sigma) \otimes t^T\Sigma] \otimes (t^T\Sigma) + m[\Sigma \otimes vec(\Sigma t \otimes \Sigma)]$$
$$+ m(\Sigma t \otimes I_{q^3})(K_{q,q} \otimes I_q)(\Sigma \otimes vec\Sigma) + mvec\{(K_{q,q} \otimes I_q)(\Sigma \otimes vec(\Sigma))\} \otimes (t^T\Sigma)$$
$$+ mvec\{vec[vec(\Sigma tt^T\Sigma)t^T\Sigma] \otimes t^T\Sigma\} \otimes (t^T\Sigma) + m\{\Sigma \otimes vec[vec(\Sigma tt^T\Sigma)t^T\Sigma]\}$$
$$+ m(\Sigma t \otimes I_{q^3})\{(\Sigma \otimes vec(\Sigma tt^T\Sigma)) + (\Sigma t \otimes I_{q^3})[(\Sigma \otimes \Sigma t) + (\Sigma t \otimes \Sigma)]\}$$
$$+ mvec\{[(\Sigma t) \otimes I_{q^3}][(\Sigma t) \otimes \Sigma]\} \otimes (t^T\Sigma) + m(I_q \otimes (\Sigma t) \otimes I_{q^3})(K_{q,q} \otimes I_q)(\Sigma \otimes vec\Sigma)$$
$$+ m(t^T\Sigma \otimes \Sigma \otimes I_{q^3})(K_{q^2,q} \otimes I_{q^2})(\Sigma \otimes vec(I_{q^2}))$$
$$+ mvec\{[(\Sigma t) \otimes I_{q^3}][\Sigma \otimes (\Sigma t)]\} \otimes (t^T\Sigma) + m(I_q \otimes \Sigma t \otimes I_{q^3})[vec\Sigma \otimes \Sigma]$$
$$+ m[\Sigma \otimes (t^T\Sigma) \otimes I_{q^3}](K_{q^2,q} \otimes I_{q^2})(\Sigma \otimes vec(I_{q^2}))$$
$$+ m\{vec[\Sigma \otimes vec(\Sigma tt^T\Sigma)] \otimes t^T\Sigma\} + m(vec\Sigma \otimes I_{q^3})\{(\Sigma \otimes \Sigma t) + (\Sigma t \otimes \Sigma)\}$$

Proof

By applying Lemma 1, I get

$$\frac{\partial}{\partial t^T} vec[m(\Sigma \otimes vec(\Sigma))] = m[vec(\Sigma \otimes vec(\Sigma)) \otimes t^T\Sigma].$$

$$\frac{\partial}{\partial t^T} vec\{mvec[vec(\Sigma)t^T\Sigma]t^T\Sigma\} = mvec\{vec[vec(\Sigma)t^T\Sigma]t^T\Sigma\} \otimes (t^T\Sigma)$$

$$+m(\Sigma \otimes vec[vec(\Sigma)t^T\Sigma]) + m(\Sigma t \otimes I_{q^3})\frac{\partial}{\partial t^T} vec[vec(\Sigma)t^T\Sigma],$$

where $\dfrac{\partial}{\partial t^T} vec[vec(\Sigma)t^T\Sigma] = (\Sigma \otimes vec(\Sigma))$.

The first term is by applying Lemma 1 again. The second term is obtained by taking derivative with respect to the t^T of the $t^T\Sigma$ at the end. The third term is by taking derivative with respect to the t^T of the $t^T\Sigma$ right after the vec(Σ).

$$\frac{\partial}{\partial t^T} vec\{m[vec(\Sigma \otimes \Sigma t) \otimes t^T\Sigma]\} = m\{vec[vec(\Sigma \otimes \Sigma t) \otimes t^T\Sigma] \otimes t^T\Sigma\} + m(\Sigma \otimes vec(\Sigma \otimes \Sigma t))$$
$$+ m(\Sigma t \otimes I_{q^2}) \frac{\partial}{\partial t^T} vec(\Sigma \otimes \Sigma t)$$
with $\frac{\partial}{\partial t^T} vec(\Sigma \otimes \Sigma t) = (I_q \otimes K_{1,q} \otimes I_q)(vec\Sigma \otimes I_q)\Sigma = (vec\Sigma \otimes \Sigma)$

The first term is obtained by taking the derivative of m; the second by taking the derivative of the $t^T\Sigma$ outside the *vec* function, and the last one inside the *vec* function.

$$\frac{\partial}{\partial t^T} vec\{m[vec(\Sigma) \otimes \Sigma]\} = m(vec[vec(\Sigma) \otimes \Sigma] \otimes t^T\Sigma), \text{ by taking derivative of m.}$$

$$\frac{\partial}{\partial t^T} vec\{m[vec(\Sigma t \otimes \Sigma) \otimes t^T\Sigma]\} = mvec[vec(\Sigma t \otimes \Sigma) \otimes t^T\Sigma] \otimes (t^T\Sigma) + m[\Sigma \otimes vec(\Sigma t \otimes \Sigma)]$$
$$+ m(\Sigma t \otimes I_{q^2}) \frac{\partial}{\partial t^T} vec(\Sigma t \otimes \Sigma)$$
with $\frac{\partial}{\partial t^T} vec(\Sigma t \otimes \Sigma) = (I_1 \otimes K_{q,q} \otimes I_q)(I_q \otimes vec\Sigma)\Sigma = (K_{q,q} \otimes I_q)(\Sigma \otimes vec\Sigma)$.

The first term is by taking derivative of m, the second of $t^T\Sigma$ outside the inner *vec* function, and the last one inside the inner *vec* function.

$$\frac{\partial}{\partial t^T} vecm\{(K_{q,q} \otimes I_q)(\Sigma \otimes vec(\Sigma))\} = mvec\{(K_{q,q} \otimes I_q)(\Sigma \otimes vec(\Sigma))\} \otimes (t^T\Sigma) \text{ by}$$

taking derivative of m.

$$\frac{\partial}{\partial t^T} vec\{m\{vec[vec(\Sigma\ t\ t^T\Sigma)t^T\Sigma]\otimes t^T\Sigma\}\} = mvec\{vec[vec(\Sigma\ t\ t^T\Sigma)t^T\Sigma]\otimes t^T\Sigma\}\otimes(t^T\Sigma)$$

$$+\{\Sigma\otimes vec[vec(\Sigma\ t\ t^T\Sigma)t^T\Sigma]\}+(\Sigma t\otimes I_{q^3})\frac{\partial}{\partial t^T} vec[vec(\Sigma\ t\ t^T\Sigma)t^T\Sigma]$$

$$\frac{\partial}{\partial t^T} vec[vec(\Sigma\ t\ t^T\Sigma)t^T\Sigma] = (\Sigma\otimes vec(\Sigma\ t\ t^T\Sigma))+(\Sigma t\otimes I_{q^2})\frac{\partial}{\partial t^T} vec(\Sigma\ t\ t^T\Sigma)$$

$$= (\Sigma\otimes vec(\Sigma\ t\ t^T\Sigma))+(\Sigma t\otimes I_{q^2})[(\Sigma\otimes\Sigma t)+(\Sigma t\otimes\Sigma)]$$

Therefore,

$$\frac{\partial}{\partial t^T} vec\{m\{vec[vec(\Sigma\ t\ t^T\Sigma)t^T\Sigma]\otimes t^T\Sigma\}\}$$

$$= mvec\{vec[vec(\Sigma\ t\ t^T\Sigma)t^T\Sigma]\otimes t^T\Sigma\}\otimes(t^T\Sigma)+m\{\Sigma\otimes vec[vec(\Sigma\ t\ t^T\Sigma)t^T\Sigma]\}$$

$$+m(\Sigma t\otimes I_{q^3})\{(\Sigma\otimes vec(\Sigma\ t\ t^T\Sigma))+(\Sigma t\otimes I_{q^2})[(\Sigma\otimes\Sigma t)+(\Sigma t\otimes\Sigma)]\}.$$

$$\frac{\partial}{\partial t^T} vec\{m[(\Sigma\ t)\otimes I_{q^2}][(\Sigma\ t)\otimes\Sigma]\} = mvec\{[(\Sigma\ t)\otimes I_{q^2}][(\Sigma\ t)\otimes\Sigma]\}\otimes(t^T\Sigma)$$

$$+m(I_q\otimes(\Sigma t)\otimes I_{q^2})\frac{\partial}{\partial t^T} vec[(\Sigma\ t)\otimes\Sigma]+m(t^T\Sigma\otimes\Sigma\otimes I_{q^3})\frac{\partial}{\partial t^T} vec[(\Sigma t)\otimes I_{q^2}],$$

where $\frac{\partial}{\partial t^T} vec[(\Sigma\ t)\otimes\Sigma] = (I_1\otimes K_{q,q}\otimes I_q)(I_q\otimes vec\Sigma)\Sigma = (K_{q,q}\otimes I_q)(\Sigma\otimes vec\Sigma)$,

and $\frac{\partial}{\partial t^T} vec[(\Sigma t)\otimes I_{q^2}] = (I_1\otimes K_{q^2,q}\otimes I_{q^2})(I_q\otimes vec(I_{q^2}))\Sigma$

$$= (K_{q^2,q}\otimes I_{q^2})(\Sigma\otimes vec(I_{q^2})).$$

So, $\frac{\partial}{\partial t^T} vec\{m[(\Sigma t)\otimes I_{q^2}][(\Sigma t)\otimes\Sigma]\} = mvec\{[(\Sigma t)\otimes I_{q^2}][(\Sigma t)\otimes\Sigma]\}\otimes(t^T\Sigma)$

$$+m(I_q\otimes(\Sigma t)\otimes I_{q^2})(K_{q,q}\otimes I_q)(\Sigma\otimes vec\Sigma)$$

$$+m(t^T\Sigma\otimes\Sigma\otimes I_{q^3})(K_{q^2,q}\otimes I_{q^2})(\Sigma\otimes vec(I_{q^2})).$$

The next term,

$$\frac{\partial}{\partial t^T} vec\{m[(\Sigma t)\otimes I_{q^2}][\Sigma\otimes(\Sigma\, t)]\} = mvec\{[(\Sigma t)\otimes I_{q^2}][\Sigma\otimes(\Sigma\, t)]\}\otimes(t^T\Sigma)$$

$$+m(I_q\otimes\Sigma t\otimes I_{q^2})\frac{\partial}{\partial t^T} vec[\Sigma\otimes(\Sigma\, t)]+m[\Sigma\otimes(t^T\Sigma)\otimes I_{q^3}]\frac{\partial}{\partial t^T} vec[(\Sigma t)\otimes I_{q^2}],$$

where $\frac{\partial}{\partial t^T} vec[\Sigma\otimes(\Sigma t)] = (I_q\otimes K_{1,q}\otimes I_q)(vec\Sigma\otimes I_q)\Sigma = (vec\Sigma\otimes\Sigma)$.

Therefore,

$$\frac{\partial}{\partial t^T} vec\{m[(\Sigma t)\otimes I_{q^2}][\Sigma\otimes(\Sigma t)]\} = mvec\{[(\Sigma t)\otimes I_{q^2}][\Sigma\otimes(\Sigma\, t)]\}\otimes(t^T\Sigma)$$
$$+m(I_q\otimes\Sigma t\otimes I_{q^2})[vec\Sigma\otimes\Sigma]+m[\Sigma\otimes(t^T\Sigma)\otimes I_{q^3}](K_{q^2,q}\otimes I_{q^2})(\Sigma\otimes vec(I_{q^2})).$$

The last term,

$$\frac{\partial}{\partial t^T} vec\{m[\Sigma\otimes vec(\Sigma tt^T\Sigma)]\} = m\{vec[\Sigma\otimes vec(\Sigma tt^T\Sigma)]\otimes t^T\Sigma\}$$
$$+ m(I_q\otimes K_{1,q}\otimes I_{q^2})(vec\Sigma\otimes I_{q^2})\frac{\partial}{\partial t^T} vec(\Sigma tt^T\Sigma)$$
$$= m\{vec[\Sigma\otimes vec(\Sigma tt^T\Sigma)]\otimes t^T\Sigma\}+m(vec\Sigma\otimes I_{q^2})\{(\Sigma\otimes\Sigma t)(\Sigma t\otimes\Sigma)\}$$

The first term of the above equation is obtained by taking the derivative of *m*; the second by the derivative of the *vec* function.

By setting t of the fifth derivative to 0, the fifth moment of a multivariate normal distribution is 0.

Sixth Derivative

Theorem 6 The sixth derivative is

$$\frac{\partial vec\dfrac{\partial vec\dfrac{\partial vec\dfrac{\partial vec\dfrac{\partial vec\dfrac{\partial^2 m}{\partial t^T \partial t}}{\partial t^T}}{\partial t^T}}{\partial t^T}}{\partial t^T}}{\partial t^T} = mvec\{[vec(\Sigma \otimes vec(\Sigma)) \otimes t^T\Sigma]\} \otimes (t^T\Sigma)$$

$$+ m(\Sigma \otimes vec(\Sigma \otimes vec(\Sigma))) + mvec\{vec\{vec[vec(\Sigma)t^T\Sigma]t^T\Sigma\} \otimes (t^T\Sigma)\} \otimes (t^T\Sigma)$$
$$+ m(\Sigma \otimes vec\{vec[vec(\Sigma)t^T\Sigma]t^T\Sigma\})$$
$$+ m(\Sigma t \otimes I_{q^4})\{(\Sigma \otimes vec[vec(\Sigma)t^T\Sigma]) + ((\Sigma t) \otimes I_{q^3})(\Sigma \otimes vec(\Sigma))\}$$
$$+ mvec(\Sigma \otimes vec(vec(\Sigma)t^T\Sigma)) \otimes (t^T\Sigma) + m(vec\Sigma \otimes I_{q^3})(\Sigma \otimes vec(\Sigma))$$
$$+ mvec[(\Sigma t \otimes I_{q^3})(\Sigma \otimes vec(\Sigma))] \otimes (t^T\Sigma)$$
$$+ m[(\Sigma \otimes (vec\Sigma)^T \otimes I_{q^4}](K_{q^3,q} \otimes I_{q^3})(\Sigma \otimes vec(I_{q^3}))$$
$$+ mvec\{\{vec[vec(\Sigma \otimes \Sigma t) \otimes t^T\Sigma] \otimes t^T\Sigma\}\} \otimes (t^T\Sigma) + m\{\Sigma \otimes vec[vec(\Sigma \otimes \Sigma t) \otimes t^T\Sigma]\}$$
$$+ m(\Sigma t \otimes I_{q^4})\{[\Sigma \otimes vec(\Sigma \otimes \Sigma t)] + [\Sigma t \otimes I_{q^3}](vec\Sigma \otimes \Sigma)\}$$
$$+ mvec\{(\Sigma \otimes vec(\Sigma \otimes \Sigma t))\} \otimes (t^T\Sigma) + m(vec\Sigma \otimes I_{q^3})(vec\Sigma \otimes \Sigma)$$
$$+ mvec[(\Sigma t \otimes I_{q^3})(vec\Sigma \otimes \Sigma)] \otimes (t^T\Sigma) + m((vec\Sigma)^T \otimes \Sigma \otimes I_{q^4})(K_{q^3,q} \otimes I_{q^3})(\Sigma \otimes vec(I_{q^3}))$$
$$+ mvec(vec[vec(\Sigma) \otimes \Sigma] \otimes t^T\Sigma) \otimes (t^T\Sigma)$$
$$+ m(\Sigma \otimes vec[vec(\Sigma) \otimes \Sigma]) + mvec\{vec[vec(\Sigma t \otimes \Sigma) \otimes t^T\Sigma] \otimes (t^T\Sigma)\} \otimes (t^T\Sigma)$$
$$+ m\{\Sigma \otimes vec[vec(\Sigma t \otimes \Sigma) \otimes t^T\Sigma]\}$$
$$+ m[(\Sigma t) \otimes I_{q^4}]\{[\Sigma \otimes vec(\Sigma t \otimes \Sigma)] + ((\Sigma t) \otimes I_{q^3})(K_{q,q} \otimes I_q)(\Sigma \otimes vec\Sigma)\}$$
$$+ mvec\{[\Sigma \otimes vec(\Sigma t \otimes \Sigma)]\} \otimes (t^T\Sigma) + m(vec\Sigma \otimes I_{q^3})(K_{q,q} \otimes I_q)(\Sigma \otimes vec\Sigma)$$
$$+ mvec[(\Sigma t \otimes I_{q^3})(K_{q,q} \otimes I_q)(\Sigma \otimes vec\Sigma)] \otimes (t^T\Sigma)$$
$$+ m[(\Sigma \otimes (vec\Sigma)^T)(K_{q,q} \otimes I_q) \otimes I_{q^4}](K_{q^3,q} \otimes I_{q^3})(\Sigma \otimes vec(I_{q^3}))$$
$$+ mvec[vec\{(K_{q,q} \otimes I_q)(\Sigma \otimes vec(\Sigma))\} \otimes (t^T\Sigma)] \otimes (t^T\Sigma)$$
$$+ m[\Sigma \otimes vec\{(K_{q,q} \otimes I_q)(\Sigma \otimes vec(\Sigma))\}]$$
$$+ mvec\{vec\{vec[vec(\Sigma tt^T\Sigma)t^T\Sigma] \otimes t^T\Sigma\} \otimes (t^T\Sigma)\} \otimes (t^T\Sigma)$$
$$+ m(\Sigma \otimes vec\{vec[vec(\Sigma tt^T\Sigma)t^T\Sigma] \otimes t^T\Sigma\} + m(\Sigma t \otimes I_{q^4})\langle\{\Sigma \otimes vec[vec(\Sigma tt^T\Sigma)t^T\Sigma]\}$$
$$+ ((\Sigma t) \otimes I_{q^3})\{[\Sigma \otimes vec(\Sigma tt^T\Sigma)] + ((\Sigma t) \otimes I_{q^3})[(\Sigma \otimes (\Sigma t)) + (\Sigma t \otimes \Sigma)]\}\rangle$$

$$+m\mathrm{vec}\{\Sigma\otimes\mathrm{vec}[\mathrm{vec}(\Sigma tt^T\Sigma)t^T\Sigma]\}\otimes(t^T\Sigma)$$

$$+(\mathrm{vec}\Sigma\otimes I_{q^3})\{[\Sigma\otimes\mathrm{vec}(\Sigma tt^T\Sigma)]+((\Sigma t)\otimes I_{q^3})[(\Sigma\otimes(\Sigma t))+(\Sigma t\otimes\Sigma)]\}$$

$$+m\mathrm{vec}[(\Sigma t\otimes I_{q^3})(\Sigma\otimes\mathrm{vec}(\Sigma tt^T\Sigma))]\otimes(t^T\Sigma)$$

$$+m(I_q\otimes\Sigma t\otimes I_{q^3})(\mathrm{vec}\Sigma\otimes I_{q^3})[(\Sigma\otimes\Sigma t)+(\Sigma t\otimes\Sigma)]$$

$$+m[\Sigma\otimes(\mathrm{vec}(\Sigma tt^T\Sigma))^T\otimes I_{q^4}](K_{q^3,q}\otimes I_{q^3})(\Sigma\otimes\mathrm{vec}(I_{q^3})$$

$$+m\mathrm{vec}[(\Sigma t\otimes I_{q^3})(\Sigma t\otimes I_{q^3})(\Sigma\otimes\Sigma t)]\otimes(t^T\Sigma)+m(I_q\otimes[(\Sigma t\otimes I_{q^3})(\Sigma t\otimes I_{q^3})])(\mathrm{vec}\Sigma\otimes\Sigma)$$

$$+m[(\Sigma\otimes t^T\Sigma)\otimes(\Sigma t\otimes I_{q^3})](K_{q^2,q}\otimes I_{q^2})(\Sigma\otimes\mathrm{vec}(I_{q^2}))$$

$$+m[(\Sigma\otimes t^T\Sigma)(t^T\Sigma\otimes I_{q^2})\otimes I_{q^4}](K_{q^3,q}\otimes I_{q^3})(\Sigma\otimes\mathrm{vec}(I_{q^3}))$$

$$+m\mathrm{vec}[(\Sigma t\otimes I_{q^3})(\Sigma t\otimes I_{q^2})(\Sigma t\otimes\Sigma)]\otimes(t^T\Sigma)$$

$$+m(I_q\otimes[(\Sigma t\otimes I_{q^3})(\Sigma t\otimes I_{q^2})])(K_{q,q}\otimes I_q)(\Sigma\otimes\mathrm{vec}\Sigma)$$

$$+m[((t^T\Sigma)\otimes\Sigma)\otimes(\Sigma t\otimes I_{q^3})](K_{q^2,q}\otimes I_{q^2})(\Sigma\otimes\mathrm{vec}(I_{q^2}))$$

$$+m[(t^T\Sigma\otimes\Sigma)(t^T\Sigma\otimes I_{q^2})\otimes I_{q^4}](K_{q^3,q}\otimes I_{q^3})(\Sigma\otimes\mathrm{vec}(I_{q^3}))$$

$$+m\mathrm{vec}\{\mathrm{vec}\{[(\Sigma t)\otimes I_{q^2}][(\Sigma t)\otimes\Sigma]\}\otimes(t^T\Sigma)\}\otimes(t^T\Sigma)$$

$$+m(\Sigma\otimes\mathrm{vec}\{[(\Sigma t)\otimes I_{q^2}][(\Sigma t)\otimes\Sigma]\})$$

$$+m[(\Sigma t)\otimes I_{q^4}]\{(I_q\otimes(\Sigma t)\otimes I_{q^2})(K_{q,q}\otimes I_q)(\Sigma\otimes\mathrm{vec}\Sigma)$$

$$+((t^T\Sigma)\otimes\Sigma\otimes I_{q^3})(K_{q^2,q}\otimes I_{q^2})(\Sigma\otimes\mathrm{vec}(I_{q^2}))$$

$$+m\mathrm{vec}[(I_q\otimes(\Sigma t)\otimes I_{q^2})(K_{q,q}\otimes I_q)(\Sigma\otimes\mathrm{vec}\Sigma)]\otimes(t^T\Sigma)$$

$$+m(\Sigma\otimes(\mathrm{vec}\Sigma)^T)(K_{q,q}\otimes I_q)\otimes I_{q^4})(I_q\otimes K_{q^2,q}\otimes I_{q^3})(\mathrm{vec}(I_q)\otimes I_{q^3})(K_{q^2,q}\otimes I_{q^2})\times$$

$$(\Sigma\otimes\mathrm{vec}(I_{q^2}))+m\mathrm{vec}[((t^T\Sigma)\otimes\Sigma\otimes I_{q^3})(K_{q^2,q}\otimes I_{q^3})(\Sigma\otimes\mathrm{vec}(I_{q^3}))]\otimes(t^T\Sigma)$$

$$+m[(\Sigma\otimes(\mathrm{vec}(I_{q^2}))^T)(K_{q,q^2}\otimes I_{q^2})\otimes I_{q^4}][\Sigma\otimes\mathrm{vec}(\Sigma\otimes I_{q^3})]$$

$$+m\mathrm{vec}\{\mathrm{vec}\{[(\Sigma t)\otimes I_{q^2}][\Sigma\otimes(\Sigma t)]\}\otimes(t^T\Sigma)\}\otimes(t^T\Sigma)$$

$$+m(\Sigma\otimes\mathrm{vec}\{[(\Sigma t)\otimes I_{q^2}][\Sigma\otimes(\Sigma t)]\})$$

$$+m((\Sigma t)\otimes I_{q^4})\{(I_q\otimes[(\Sigma t)\otimes I_{q^2}])(\mathrm{vec}\Sigma\otimes\Sigma)$$

$$+([\Sigma\otimes(t^T\Sigma)\otimes I_{q^3}](K_{q^2,q}\otimes I_{q^2})(\Sigma\otimes\mathrm{vec}(I_{q^2}))\}$$

$$+m\mathrm{vec}[(I_q\otimes\Sigma t\otimes I_{q^2})(\mathrm{vec}\Sigma\otimes\Sigma)]\otimes(t^T\Sigma)$$

$$+m((\mathrm{vec}\Sigma)^T\otimes\Sigma\otimes I_{q^4})(I_q\otimes K_{q^3,q}\otimes I_{q^3})(\mathrm{vec}(I_q)\otimes I_{q^3})(K_{q^3,q}\otimes I_{q^3})(\Sigma\otimes\mathrm{vec}(I_{q^3})$$

$$+m\mathrm{vec}\{[\Sigma\otimes(t^T\Sigma)\otimes I_{q^3}](K_{q^2,q}\otimes I_{q^2})(\Sigma\otimes\mathrm{vec}(I_{q^2}))\}\otimes(t^T\Sigma)$$

$$+m\{[(\Sigma\otimes(\mathrm{vec}(I_{q^2}))^T)(K_{q,q^2}\otimes I_{q^2})]\otimes I_{q^4}\}(I_{q^2}\otimes K_{q^3,q}\otimes I_{q^3})(I_{q^3}\otimes\mathrm{vec}(I_{q^3}))(\mathrm{vec}\Sigma\otimes\Sigma)$$

$$+m\mathrm{vec}\{\mathrm{vec}[\Sigma \otimes \mathrm{vec}(\Sigma t t^T \Sigma)] \otimes (t^T \Sigma)\} \otimes (t^T \Sigma)$$

$$+m\{\Sigma \otimes \mathrm{vec}[\Sigma \otimes \mathrm{vec}(\Sigma t t^T \Sigma)]\} + m(\Sigma t \otimes I_{q^4})(\mathrm{vec}\Sigma \otimes I_{q^2})[(\Sigma \otimes \Sigma t) + (\Sigma t \otimes \Sigma)]$$

$$+m\mathrm{vec}\{(\mathrm{vec}\Sigma \otimes I_{q^2})[(\Sigma \otimes \Sigma t) + (\Sigma t \otimes \Sigma)]\} \otimes (t^T \Sigma)$$

$$+m(I_q \otimes (\mathrm{vec}\Sigma \otimes I_{q^2}))[(\mathrm{vec}\Sigma \otimes \Sigma) + (K_{qq} \otimes I_q)(\Sigma \otimes \mathrm{vec}\Sigma)]$$

Proof

$$\frac{\partial}{\partial t^T} \mathrm{vec}\{m[\mathrm{vec}(\Sigma \otimes \mathrm{vec}(\Sigma)) \otimes t^T \Sigma]\} = m\mathrm{vec}\{[\mathrm{vec}(\Sigma \otimes \mathrm{vec}(\Sigma)) \otimes t^T \Sigma]\} \otimes (t^T \Sigma)$$

$$+m(\Sigma \otimes \mathrm{vec}(\Sigma \otimes \mathrm{vec}(\Sigma)))$$

$$\frac{\partial}{\partial t^T} \mathrm{vec}\{m\mathrm{vec}\{\mathrm{vec}[\mathrm{vec}(\Sigma) t^T \Sigma] t^T \Sigma\} \otimes (t^T \Sigma)\}$$

$$= m\mathrm{vec}\{\mathrm{vec}\{\mathrm{vec}[\mathrm{vec}(\Sigma) t^T \Sigma] t^T \Sigma\} \otimes (t^T \Sigma)\} \otimes (t^T \Sigma)$$

$$+m(\Sigma \otimes \mathrm{vec}\{\mathrm{vec}[\mathrm{vec}(\Sigma) t^T \Sigma] t^T \Sigma\}) + m(\Sigma t \otimes I_{q^4}) \frac{\partial}{\partial t^T} \mathrm{vec}\{\mathrm{vec}[\mathrm{vec}(\Sigma) t^T \Sigma] t^T \Sigma\}$$

where $\frac{\partial}{\partial t^T} \mathrm{vec}\{\mathrm{vec}[\mathrm{vec}(\Sigma) t^T \Sigma] t^T \Sigma\}$

$$= (\Sigma \otimes \mathrm{vec}[\mathrm{vec}(\Sigma) t^T \Sigma]) + ((\Sigma t) \otimes I_{q^3}) \frac{\partial}{\partial t^T} \mathrm{vec}[\mathrm{vec}(\Sigma) t^T \Sigma]$$

with $\frac{\partial}{\partial t^T} \mathrm{vec}[\mathrm{vec}(\Sigma) t^T \Sigma] = (\Sigma \otimes \mathrm{vec}(\Sigma))$.

Therefore,

$$\frac{\partial}{\partial t^T} \mathrm{vec}\{\mathrm{vec}[\mathrm{vec}(\Sigma) t^T \Sigma] t^T \Sigma\} = (\Sigma \otimes \mathrm{vec}[\mathrm{vec}(\Sigma) t^T \Sigma]) + ((\Sigma t) \otimes I_{q^3})[\Sigma \otimes \mathrm{vec}(\Sigma)].$$

$$\frac{\partial}{\partial t^T} \mathrm{vec}\{m\mathrm{vec}\{\mathrm{vec}[\mathrm{vec}(\Sigma) t^T \Sigma] t^T \Sigma\} \otimes (t^T \Sigma)\}$$

$$= m\mathrm{vec}\{\mathrm{vec}\{\mathrm{vec}[\mathrm{vec}(\Sigma) t^T \Sigma] t^T \Sigma\} \otimes (t^T \Sigma)\} \otimes (t^T \Sigma) + m(\Sigma \otimes \mathrm{vec}\{\mathrm{vec}[\mathrm{vec}(\Sigma) t^T \Sigma] t^T \Sigma\})$$

$$+m(\Sigma t \otimes I_{q^4})\{(\Sigma \otimes \mathrm{vec}[\mathrm{vec}(\Sigma) t^T \Sigma]) + ((\Sigma t) \otimes I_{q^3})(\Sigma \otimes \mathrm{vec}(\Sigma))\}$$

$$\frac{\partial}{\partial t^T} vec\{m(\Sigma \otimes vec(vec(\Sigma)t^T\Sigma))\} = mvec(\Sigma \otimes vec(vec(\Sigma)t^T\Sigma)) \otimes (t^T\Sigma)$$

$$+m(I_q \otimes K_{1,q} \otimes I_{q^3})(vec\Sigma \otimes I_{q^3})\frac{\partial}{\partial t^T} vec(vec(\Sigma)t^T\Sigma))$$

while $\frac{\partial}{\partial t^T} vec[vec(\Sigma)t^T\Sigma] = (\Sigma \otimes vec(\Sigma))$, whereas $K_{1,q} = I_q$ and can be ignored.

Therefore, $\frac{\partial}{\partial t^T} vec\{m(\Sigma \otimes vec(vec(\Sigma)t^T\Sigma))\} = mvec(\Sigma \otimes vec(vec(\Sigma)t^T\Sigma)) \otimes (t^T\Sigma)$

$$+ m(vec\Sigma \otimes I_{q^3})(\Sigma \otimes vec(\Sigma)).$$

$$\frac{\partial}{\partial t^T} vec[m(\Sigma t \otimes I_{q^3})(\Sigma \otimes vec(\Sigma))] = mvec[(\Sigma t \otimes I_{q^3})(\Sigma \otimes vec(\Sigma))] \otimes (t^T\Sigma)$$

$$+ m[(\Sigma \otimes (vec\Sigma)^T \otimes I_{q^4}]\frac{\partial}{\partial t^T} vec(\Sigma t \otimes I_{q^3})$$

with $\frac{\partial}{\partial t^T} vec(\Sigma t \otimes I_{q^3}) = (I_1 \otimes K_{q^3,q} \otimes I_{q^3})(I_q \otimes vec(I_{q^3}))\Sigma.$

$$= (K_{q^3,q} \otimes I_{q^3})(\Sigma \otimes vec(I_{q^3})).$$

So, $\frac{\partial}{\partial t^T} vec[m(\Sigma t \otimes I_{q^3})(\Sigma \otimes vec(\Sigma))]$

$$= mvec[(\Sigma t \otimes I_{q^3})(\Sigma \otimes vec(\Sigma))] \otimes (t^T\Sigma)$$
$$+m[(\Sigma \otimes (vec\Sigma)^T \otimes I_{q^4}](K_{q^3,q} \otimes I_{q^3})(\Sigma \otimes vec(I_{q^3})).$$

$$\frac{\partial}{\partial t^T} vec\{m\{vec[vec(\Sigma \otimes \Sigma t) \otimes t^T\Sigma] \otimes t^T\Sigma\}\}$$
$$= mvec\{vec[vec(\Sigma \otimes \Sigma t) \otimes t^T\Sigma] \otimes t^T\Sigma\} \otimes (t^T\Sigma)$$
$$+m\{\Sigma \otimes vec[vec(\Sigma \otimes \Sigma t) \otimes t^T\Sigma]\} + m(\Sigma t \otimes I_{q^4})\frac{\partial}{\partial t^T} vec[vec(\Sigma \otimes \Sigma t) \otimes t^T\Sigma],$$

where $\frac{\partial}{\partial t^T} vec[vec(\Sigma \otimes \Sigma t) \otimes t^T\Sigma] = [\Sigma \otimes vec(\Sigma \otimes \Sigma t)] + [\Sigma t \otimes I_{q^3}]\frac{\partial}{\partial t^T} vec(\Sigma \otimes \Sigma t),$

with $\frac{\partial}{\partial t^T} vec(\Sigma \otimes \Sigma t) = (I_q \otimes K_{1,q} \otimes I_q)(vec\Sigma \otimes I_q)\Sigma = (vec\Sigma \otimes \Sigma)$.

Therefore, $\frac{\partial}{\partial t^T} vec\{m\{vec[vec(\Sigma \otimes \Sigma t) \otimes t^T\Sigma] \otimes t^T\Sigma\}\}$

$= mvec\{\{vec[vec(\Sigma \otimes \Sigma t) \otimes t^T\Sigma] \otimes t^T\Sigma\}\} \otimes (t^T\Sigma)$

$+m\{\Sigma \otimes vec[vec(\Sigma \otimes \Sigma t) \otimes t^T\Sigma]\}$

$+m(\Sigma t \otimes I_{q^4})\{[\Sigma \otimes vec(\Sigma \otimes \Sigma t)] + [\Sigma t \otimes I_{q^3}](vec\Sigma \otimes \Sigma)\}$.

$\frac{\partial}{\partial t^T} vec\{m(\Sigma \otimes vec(\Sigma \otimes \Sigma t))\} = mvec(\Sigma \otimes vec(\Sigma \otimes \Sigma t)) \otimes (t^T\Sigma)$

$+ m(I_q \otimes K_{1,q} \otimes I_{q^3})(vec\Sigma \otimes I_{q^3}) \frac{\partial}{\partial t^T} vec(\Sigma \otimes \Sigma t)$

$= mvec\{(\Sigma \otimes vec(\Sigma \otimes \Sigma t))\} \otimes (t^T\Sigma) + m(vec\Sigma \otimes I_{q^3})(vec\Sigma \otimes \Sigma)$

$\frac{\partial}{\partial t^T} vec[m(\Sigma t \otimes I_{q^3})(vec\Sigma \otimes \Sigma)] = mvec[(\Sigma t \otimes I_{q^3})(vec\Sigma \otimes \Sigma)] \otimes (t^T\Sigma)$

$+m((vec\Sigma)^T \otimes \Sigma \otimes I_{q^4}) \frac{\partial}{\partial t^T} vec(\Sigma t \otimes I_{q^3})$

$= mvec[(\Sigma t \otimes I_{q^3})(vec\Sigma \otimes \Sigma)] \otimes (t^T\Sigma)$

$+m((vec\Sigma)^T \otimes \Sigma \otimes I_{q^4})(K_{q^3,q} \otimes I_{q^3})(\Sigma \otimes vec(I_{q^3}))$

$\frac{\partial}{\partial t^T} vec\{m(vec[vec(\Sigma) \otimes \Sigma] \otimes t^T\Sigma)\} = mvec(vec[vec(\Sigma) \otimes \Sigma] \otimes t^T\Sigma) \otimes (t^T\Sigma)$

$+ m(\Sigma \otimes vec[vec(\Sigma) \otimes \Sigma])$

$\frac{\partial}{\partial t^T} vec\{mvec[vec(\Sigma t \otimes \Sigma) \otimes t^T\Sigma] \otimes (t^T\Sigma)\}$

$= mvec\{vec[vec(\Sigma t \otimes \Sigma) \otimes t^T\Sigma] \otimes (t^T\Sigma)\} \otimes (t^T\Sigma) + m\{\Sigma \otimes vec[vec(\Sigma t \otimes \Sigma) \otimes t^T\Sigma]\}$

$+m[(\Sigma t) \otimes I_{q^4}] \frac{\partial}{\partial t^T} vec[vec(\Sigma t \otimes \Sigma) \otimes t^T\Sigma]$,

with $\frac{\partial}{\partial t^T} vec[vec(\Sigma t \otimes \Sigma) \otimes t^T\Sigma]$

$$= [\Sigma \otimes vec(\Sigma t \otimes \Sigma)] + ((\Sigma t) \otimes I_{q^3}) \frac{\partial}{\partial t^T} vec(\Sigma t \otimes \Sigma)$$
$$= [\Sigma \otimes vec(\Sigma t \otimes \Sigma)] + ((\Sigma t) \otimes I_{q^3})(K_{q,q} \otimes I_q)(\Sigma \otimes vec\Sigma)$$

Therefore,

$$\frac{\partial}{\partial t^T} vec\{mvec[vec(\Sigma t \otimes \Sigma) \otimes t^T\Sigma] \otimes (t^T\Sigma)\}$$
$$= mvec\{vec[vec(\Sigma t \otimes \Sigma) \otimes t^T\Sigma] \otimes (t^T\Sigma)\} \otimes (t^T\Sigma) + m\{\Sigma \otimes vec[vec(\Sigma t \otimes \Sigma) \otimes t^T\Sigma]\}$$
$$+ m[(\Sigma t) \otimes I_{q^4}]\{[\Sigma \otimes vec(\Sigma t \otimes \Sigma)] + ((\Sigma t) \otimes I_{q^3})(K_{q,q} \otimes I_q)(\Sigma \otimes vec\Sigma)\}.$$

$$\frac{\partial}{\partial t^T} vec\{m[\Sigma \otimes vec(\Sigma t \otimes \Sigma)]\} = mvec\{[\Sigma \otimes vec(\Sigma t \otimes \Sigma)]\} \otimes (t^T\Sigma)$$
$$+ m(I_q \otimes K_{1,q} \otimes I_{q^3})(vec\Sigma \otimes I_{q^3}) \frac{\partial}{\partial t^T} vec(\Sigma t \otimes \Sigma)$$
$$= mvec\{[\Sigma \otimes vec(\Sigma t \otimes \Sigma)]\} \otimes (t^T\Sigma) + m(vec\Sigma \otimes I_{q^3})(K_{q,q} \otimes I_q)(\Sigma \otimes vec\Sigma)$$

$$\frac{\partial}{\partial t^T} vec[m(\Sigma t \otimes I_{q^3})(K_{q,q} \otimes I_q)(\Sigma \otimes vec\Sigma)]$$
$$= mvec[(\Sigma t \otimes I_{q^3})(K_{q,q} \otimes I_q)(\Sigma \otimes vec\Sigma)] \otimes t^T\Sigma$$
$$+ m[(\Sigma \otimes (vec\Sigma)^T)(K_{q,q} \otimes I_q) \otimes I_{q^4}] \frac{\partial}{\partial t^T} (\Sigma t \otimes I_{q^3})$$
$$= mvec[(\Sigma t \otimes I_{q^3})(K_{q,q} \otimes I_q)(\Sigma \otimes vec\Sigma)] \otimes (t^T\Sigma)$$
$$+ m[(\Sigma \otimes (vec\Sigma)^T)(K_{q,q} \otimes I_q) \otimes I_{q^4}](K_{q^3,q} \otimes I_{q^3})(\Sigma \otimes vec(I_{q^3}))$$

$$\frac{\partial}{\partial t^T} vec[mvec\{(K_{q,q} \otimes I_q)(\Sigma \otimes vec(\Sigma))\} \otimes (t^T\Sigma)]$$
$$= mvec[vec\{(K_{q,q} \otimes I_q)(\Sigma \otimes vec(\Sigma))\} \otimes (t^T\Sigma)] \otimes (t^T\Sigma)$$
$$+ m[\Sigma \otimes vec\{(K_{q,q} \otimes I_q)(\Sigma \otimes vec(\Sigma))\}]$$

$$\frac{\partial}{\partial t^T} vec\{mvec\{vec[vec(\Sigma t t^T\Sigma)t^T\Sigma] \otimes t^T\Sigma\} \otimes (t^T\Sigma)\}$$
$$= mvec\{vec\{vec[vec(\Sigma t t^T\Sigma)t^T\Sigma] \otimes t^T\Sigma\} \otimes (t^T\Sigma)\} \otimes (t^T\Sigma)$$
$$+ m(\Sigma \otimes vec\{vec[vec(\Sigma t t^T\Sigma)t^T\Sigma] \otimes t^T\Sigma\}$$
$$+ m(\Sigma t \otimes I_{q^4}) \frac{\partial}{\partial t^T} vec\{vec[vec(\Sigma t t^T\Sigma)t^T\Sigma] \otimes t^T\Sigma\},$$

with $\frac{\partial}{\partial t^T} vec\{vec[vec(\Sigma t t^T \Sigma) t^T \Sigma] \otimes t^T \Sigma\}$

$$= \{\Sigma \otimes vec[vec(\Sigma t t^T \Sigma) t^T \Sigma]\} + ((\Sigma t) \otimes I_{q^3}) \frac{\partial}{\partial t^T} vec[vec(\Sigma t t^T \Sigma) t^T \Sigma]$$
$$= \{\Sigma \otimes vec[vec(\Sigma t t^T \Sigma) t^T \Sigma]\} + ((\Sigma t) \otimes I_{q^3}) \{[\Sigma \otimes vec(\Sigma t t^T \Sigma)]$$
$$+ [((\Sigma t) \otimes I_{q^2})[(\Sigma \otimes (\Sigma t)) + (\Sigma t \otimes \Sigma)]\}$$

Therefore, $\frac{\partial}{\partial t^T} vec\{mvec\{vec[vec(\Sigma t t^T \Sigma) t^T \Sigma] \otimes t^T \Sigma\} \otimes (t^T \Sigma)\}$

$$= mvec\{vec\{vec[vec(\Sigma t t^T \Sigma) t^T \Sigma] \otimes t^T \Sigma\} \otimes (t^T \Sigma)\} \otimes (t^T \Sigma)$$
$$+ m(\Sigma \otimes vec\{vec[vec(\Sigma t t^T \Sigma) t^T \Sigma] \otimes t^T \Sigma\} + m(\Sigma t \otimes I_{q^4}) \langle \{\Sigma \otimes vec[vec(\Sigma t t^T \Sigma) t^T \Sigma]\}$$
$$+ ((\Sigma t) \otimes I_{q^3}) \{[\Sigma \otimes vec(\Sigma t t^T \Sigma)] + ((\Sigma t) \otimes I_{q^2})[(\Sigma \otimes (\Sigma t)) + (\Sigma t \otimes \Sigma)]\} \rangle$$

$$\frac{\partial}{\partial t^T} vec(m\{\Sigma \otimes vec[vec(\Sigma t t^T \Sigma) t^T \Sigma]\}) = mvec\{\Sigma \otimes vec[vec(\Sigma t t^T \Sigma) t^T \Sigma]\} \otimes (t^T \Sigma)$$
$$+ (I_q \otimes K_{1,q} \otimes I_{q^3})(vec\Sigma \otimes I_{q^3}) \frac{\partial}{\partial t^T} vec[vec(\Sigma t t^T \Sigma) t^T \Sigma]$$
$$= mvec\{\Sigma \otimes vec[vec(\Sigma t t^T \Sigma) t^T \Sigma]\} \otimes (t^T \Sigma)$$
$$+ (vec\Sigma \otimes I_{q^3}) \{[\Sigma \otimes vec(\Sigma t t^T \Sigma)] + ((\Sigma t) \otimes I_{q^2})[(\Sigma \otimes (\Sigma t)) + (\Sigma t \otimes \Sigma)]\},$$

since $\frac{\partial}{\partial t^T} vec[vec(\Sigma t t^T \Sigma) t^T \Sigma]$

$$= \{[\Sigma \otimes vec(\Sigma t t^T \Sigma)] + [((\Sigma t) \otimes I_{q^2}) \frac{\partial}{\partial t^T} vec(\Sigma t t^T \Sigma)\}$$
$$= \{[\Sigma \otimes vec(\Sigma t t^T \Sigma)] + [((\Sigma t) \otimes I_{q^2})[(\Sigma \otimes (\Sigma t)) + (\Sigma t \otimes \Sigma)]]\}.$$

$$\frac{\partial}{\partial t^T} vec[m(\Sigma t \otimes I_{q^3})(\Sigma \otimes vec(\Sigma t t^T \Sigma))] = mvec[(\Sigma t \otimes I_{q^3})(\Sigma \otimes vec(\Sigma t t^T \Sigma))] \otimes (t^T \Sigma)$$

$$+m(I_q \otimes (\Sigma t \otimes I_{q^3})) \frac{\partial}{\partial t^T} vec(\Sigma \otimes vec(\Sigma t t^T \Sigma))$$

$$+m[\Sigma \otimes (vec(\Sigma t t^T \Sigma))^T \otimes I_{q^4}] \frac{\partial}{\partial t^T} vec(\Sigma t \otimes I_{q^3})$$

$$= mvec[(\Sigma t \otimes I_{q^3})(\Sigma \otimes vec(\Sigma t t^T \Sigma))] \otimes (t^T \Sigma)$$

$$+m(I_q \otimes \Sigma t \otimes I_{q^3})(vec\Sigma \otimes I_{q^2})[(\Sigma \otimes \Sigma t) + (\Sigma t \otimes \Sigma)]$$

$$+m[\Sigma \otimes (vec(\Sigma t t^T \Sigma))^T \otimes I_{q^4}](K_{q^3,q} \otimes I_{q^3})(\Sigma \otimes vec(I_{q^3}))$$

$$\frac{\partial}{\partial t^T} vec[m(\Sigma t \otimes I_{q^3})(\Sigma t \otimes I_{q^2})(\Sigma \otimes \Sigma t)] = mvec[(\Sigma t \otimes I_{q^3})(\Sigma t \otimes I_{q^2})(\Sigma \otimes \Sigma t)] \otimes (t^T \Sigma)$$

$$+m(I_q \otimes [(\Sigma t \otimes I_{q^3})(\Sigma t \otimes I_{q^2})]) \frac{\partial}{\partial t^T} vec(\Sigma \otimes \Sigma t)$$

$$+m[(\Sigma \otimes t^T \Sigma) \otimes (\Sigma t \otimes I_{q^3})] \frac{\partial}{\partial t^T} vec(\Sigma t \otimes I_{q^2})$$

$$+m[(\Sigma \otimes t^T \Sigma)(t^T \Sigma \otimes I_{q^2}) \otimes I_{q^4}] \frac{\partial}{\partial t^T} vec(\Sigma t \otimes I_{q^3})$$

$$= mvec[(\Sigma t \otimes I_{q^3})(\Sigma t \otimes I_{q^2})(\Sigma \otimes \Sigma t)] \otimes (t^T \Sigma)$$

$$+m(I_q \otimes [(\Sigma t \otimes I_{q^3})(\Sigma t \otimes I_{q^2})])(vec\Sigma \otimes \Sigma)$$

$$+m[(\Sigma \otimes t^T \Sigma) \otimes (\Sigma t \otimes I_{q^3})](K_{q^2,q} \otimes I_{q^2})(\Sigma \otimes vec(I_{q^2}))$$

$$+m[(\Sigma \otimes t^T \Sigma)(t^T \Sigma \otimes I_{q^2}) \otimes I_{q^4}](K_{q^3,q} \otimes I_{q^3})(\Sigma \otimes vec(I_{q^3}))$$

$$\frac{\partial}{\partial t^T} vec[m(\Sigma t \otimes I_{q^3})(\Sigma t \otimes I_{q^2})(\Sigma t \otimes \Sigma)] = mvec[(\Sigma t \otimes I_{q^3})(\Sigma t \otimes I_{q^2})(\Sigma t \otimes \Sigma)] \otimes (t^T \Sigma)$$

$$+m(I_q \otimes [(\Sigma t \otimes I_{q^3})(\Sigma t \otimes I_{q^2})])(K_{q,q} \otimes I_q)(\Sigma \otimes vec\Sigma)$$

$$+m[((t^T \Sigma) \otimes \Sigma) \otimes (\Sigma t \otimes I_{q^3})](K_{q^2,q} \otimes I_{q^2})(\Sigma \otimes vec(I_{q^2}))$$

$$+m[(t^T \Sigma \otimes \Sigma)(t^T \Sigma \otimes I_{q^2}) \otimes I_{q^4}](K_{q^3,q} \otimes I_{q^3})(\Sigma \otimes vec(I_{q^3}))$$

$$\frac{\partial}{\partial t^T} vec\{mvec\{[(\Sigma t)\otimes I_{q^2}][(\Sigma t)\otimes\Sigma]\}\otimes(t^T\Sigma)\}$$

$$= mvec\{vec\{[(\Sigma t)\otimes I_{q^2}][(\Sigma t)\otimes\Sigma]\}\otimes(t^T\Sigma)\}\otimes(t^T\Sigma)$$

$$+m(\Sigma\otimes vec\{[(\Sigma t)\otimes I_{q^2}][(\Sigma t)\otimes\Sigma]\})+m[(\Sigma t)\otimes I_{q^4}]\frac{\partial}{\partial t^T} vec\{[(\Sigma t)\otimes I_{q^2}][(\Sigma t)\otimes\Sigma]\}$$

$$= mvec\{vec\{[(\Sigma t)\otimes I_{q^2}][(\Sigma t)\otimes\Sigma]\}\otimes(t^T\Sigma)\}\otimes(t^T\Sigma)$$

$$+m(\Sigma\otimes vec\{[(\Sigma t)\otimes I_{q^2}][(\Sigma t)\otimes\Sigma]\})$$

$$+m[(\Sigma t)\otimes I_{q^4}]\{(I_q\otimes[(\Sigma t)\otimes I_{q^2}])\frac{\partial}{\partial t^T} vec[(\Sigma t)\otimes\Sigma]$$

$$+((t^T\Sigma)\otimes\Sigma)\otimes I_{q^3})\frac{\partial}{\partial t^T} vec[(\Sigma t)\otimes I_{q^2}]\}$$

$$= mvec\{vec\{[(\Sigma t)\otimes I_{q^2}][(\Sigma t)\otimes\Sigma]\}\otimes(t^T\Sigma)\}\otimes(t^T\Sigma)$$

$$+m(\Sigma\otimes vec\{[(\Sigma t)\otimes I_{q^2}][(\Sigma t)\otimes\Sigma]\})$$

$$+m[(\Sigma t)\otimes I_{q^4}]\{(I_q\otimes(\Sigma t)\otimes I_{q^2})(K_{qq}\otimes I_q)(\Sigma\otimes vec\Sigma)$$

$$+((t^T\Sigma)\otimes\Sigma\otimes I_{q^3})(K_{q^2q}\otimes I_{q^2})(\Sigma\otimes vec(I_{q^2}))$$

$$\frac{\partial}{\partial t^T} vec[m(I_q\otimes(\Sigma t)\otimes I_{q^2})(K_{qq}\otimes I_q)(\Sigma\otimes vec\Sigma)]$$

$$= mvec[(I_q\otimes(\Sigma t)\otimes I_{q^2})(K_{qq}\otimes I_q)(\Sigma\otimes vec\Sigma)]\otimes(t^T\Sigma)$$

$$+m(\Sigma\otimes(vec\Sigma)^T)(K_{qq}\otimes I_q)\otimes I_{q^4})\frac{\partial}{\partial t^T} vec(I_q\otimes(\Sigma t)\otimes I_{q^2})$$

$$= mvec[(I_q\otimes(\Sigma t)\otimes I_{q^2})(K_{qq}\otimes I_q)(\Sigma\otimes vec\Sigma)]\otimes(t^T\Sigma)$$

$$+m(\Sigma\otimes(vec\Sigma)^T)(K_{qq}\otimes I_q)\otimes I_{q^4})(I_q\otimes K_{q^2q}\otimes I_{q^3})(vec(I_q)\otimes I_{q^5})\frac{\partial}{\partial t^T} vec((\Sigma t)\otimes I_{q^2})$$

$$= mvec[(I_q\otimes(\Sigma t)\otimes I_{q^2})(K_{qq}\otimes I_q)(\Sigma\otimes vec\Sigma)]\otimes(t^T\Sigma)$$

$$+m(\Sigma\otimes(vec\Sigma)^T)(K_{qq}\otimes I_q)\otimes I_{q^4})(I_q\otimes K_{q^2q}\otimes I_{q^3})(vec(I_q)\otimes I_{q^5})(K_{q^2q}\otimes I_{q^2})\times$$

$$(\Sigma\otimes vec(I_{q^2}))$$

$$\frac{\partial}{\partial t^T} vec[m((t^T\Sigma)\otimes\Sigma\otimes I_{q^3})(K_{q^2,q}\otimes I_{q^2})(\Sigma\otimes vec(I_{q^2}))]$$
$$= mvec[((t^T\Sigma)\otimes\Sigma\otimes I_{q^3})(K_{q^2,q}\otimes I_{q^2})(\Sigma\otimes vec(I_{q^2}))]\otimes(t^T\Sigma)$$
$$+m[(\Sigma\otimes(vec(I_{q^2}))^T)(K_{q,q^2}\otimes I_{q^2})\otimes I_{q^4}]\frac{\partial}{\partial t^T} vec(t^T\Sigma\otimes\Sigma\otimes I_{q^3})$$
$$= mvec[((t^T\Sigma)\otimes\Sigma\otimes I_{q^3})(K_{q^2,q}\otimes I_{q^2})(\Sigma\otimes vec(I_{q^2}))]\otimes(t^T\Sigma)$$
$$+m[(\Sigma\otimes(vec(I_{q^2}))^T)(K_{q,q^2}\otimes I_{q^2})\otimes I_{q^4}][\Sigma\otimes vec(\Sigma\otimes I_{q^3})]$$

$$\frac{\partial}{\partial t^T} vec\{mvec\{[(\Sigma t)\otimes I_{q^2}][\Sigma\otimes(\Sigma t)]\}\otimes(t^T\Sigma)\}$$
$$= mvec\{vec\{[(\Sigma t)\otimes I_{q^2}][\Sigma\otimes(\Sigma t)]\}\otimes(t^T\Sigma)\}\otimes(t^T\Sigma)$$
$$+m(\Sigma\otimes vec\{[(\Sigma t)\otimes I_{q^2}][\Sigma\otimes(\Sigma t)]\})$$
$$+m((\Sigma t)\otimes I_{q^4})\frac{\partial}{\partial t^T} vec\{[(\Sigma t)\otimes I_{q^2}][\Sigma\otimes(\Sigma t)]\}$$
$$= mvec\{vec\{[(\Sigma t)\otimes I_{q^2}][\Sigma\otimes(\Sigma t)]\}\otimes(t^T\Sigma)\}\otimes(t^T\Sigma)$$
$$+m(\Sigma\otimes vec\{[(\Sigma t)\otimes I_{q^2}][\Sigma\otimes(\Sigma t)]\})$$
$$+m((\Sigma t)\otimes I_{q^4})\{(I_q\otimes[(\Sigma t)\otimes I_{q^2}])(vec\Sigma\otimes\Sigma)+([\Sigma\otimes(t^T\Sigma)\otimes I_{q^3}]\frac{\partial}{\partial t^T} vec[(\Sigma t)\otimes I_{q^2}]\}$$
$$= mvec\{vec\{[(\Sigma t)\otimes I_{q^2}][\Sigma\otimes(\Sigma t)]\}\otimes(t^T\Sigma)\}\otimes(t^T\Sigma)$$
$$+m(\Sigma\otimes vec\{[(\Sigma t)\otimes I_{q^2}][\Sigma\otimes(\Sigma t)]\})$$
$$+m((\Sigma t)\otimes I_{q^4})\{(I_q\otimes[(\Sigma t)\otimes I_{q^2}])(vec\Sigma\otimes\Sigma)$$
$$+([\Sigma\otimes(t^T\Sigma)\otimes I_{q^3}](K_{q^2,q}\otimes I_{q^2})(\Sigma\otimes vec(I_{q^2}))\}$$

$$\frac{\partial}{\partial t^T} vec[m(I_q\otimes\Sigma t\otimes I_{q^2})(vec\Sigma\otimes\Sigma)] = mvec[(I_q\otimes\Sigma t\otimes I_{q^2})(vec\Sigma\otimes\Sigma)]\otimes(t^T\Sigma)$$
$$+((vec\Sigma)^T\otimes\Sigma\otimes I_{q^4})\frac{\partial}{\partial t^T} vec(I_q\otimes\Sigma t\otimes I_{q^2})$$
$$= mvec[(I_q\otimes\Sigma t\otimes I_{q^2})(vec\Sigma\otimes\Sigma)]\otimes(t^T\Sigma)$$
$$+((vec\Sigma)^T\otimes\Sigma\otimes I_{q^4})(I_q\otimes K_{q^2,q}\otimes I_{q^2})(vec(I_q)\otimes I_{q^5})(K_{q^2,q}\otimes I_{q^2})(\Sigma\otimes vec(I_{q^2}))$$

$$\frac{\partial}{\partial t^T} vec\{m[\Sigma \otimes (t^T\Sigma) \otimes I_{q^3}](K_{q^2,q} \otimes I_{q^2})(\Sigma \otimes vec(I_{q^2}))\}$$
$$= mvec\{[\Sigma \otimes (t^T\Sigma) \otimes I_{q^3}](K_{q^2,q} \otimes I_{q^2})(\Sigma \otimes vec(I_{q^2}))\} \otimes (t^T\Sigma)$$
$$+ m\{[(\Sigma \otimes (vec(I_{q^2}))^T)(K_{q,q^2} \otimes I_{q^2})] \otimes I_{q^4}\} \frac{\partial}{\partial t^T} vec[\Sigma \otimes (t^T\Sigma) \otimes I_{q^3}]$$
$$= mvec\{[\Sigma \otimes (t^T\Sigma) \otimes I_{q^3}](K_{q^2,q} \otimes I_{q^2})(\Sigma \otimes vec(I_{q^2}))\} \otimes (t^T\Sigma)$$
$$+ m\{[(\Sigma \otimes (vec(I_{q^2}))^T)(K_{q,q^2} \otimes I_{q^2})] \otimes I_{q^4}\}(I_{q^2} \otimes K_{q^3,q} \otimes I_{q^3})(I_{q^3} \otimes vec(I_{q^3}))(vec\Sigma \otimes \Sigma)$$

$$\frac{\partial}{\partial t^T} vec\{mvec[\Sigma \otimes vec(\Sigma t t^T \Sigma)] \otimes t^T\Sigma\} = mvec\{vec[\Sigma \otimes vec(\Sigma t t^T \Sigma)] \otimes (t^T\Sigma)\} \otimes (t^T\Sigma)$$
$$+ m\{\Sigma \otimes vec[\Sigma \otimes vec(\Sigma t t^T \Sigma)]\} + m(\Sigma t \otimes I_{q^4}) \frac{\partial}{\partial t^T} vec[\Sigma \otimes vec(\Sigma t t^T \Sigma)]$$
$$= mvec\{vec[\Sigma \otimes vec(\Sigma t t^T \Sigma)] \otimes (t^T\Sigma)\} \otimes (t^T\Sigma)$$
$$+ m\{\Sigma \otimes vec[\Sigma \otimes vec(\Sigma t t^T \Sigma)]\} + m(\Sigma t \otimes I_{q^4})(vec\Sigma \otimes I_{q^3})[(\Sigma \otimes \Sigma t) + (\Sigma t \otimes \Sigma)]$$

$$\frac{\partial}{\partial t^T} vec\{m(vec\Sigma \otimes I_{q^2})[(\Sigma \otimes \Sigma t) + (\Sigma t \otimes \Sigma)]\}$$
$$= mvec\{(vec\Sigma \otimes I_{q^2})[(\Sigma \otimes \Sigma t) + (\Sigma t \otimes \Sigma)]\} \otimes (t^T\Sigma)$$
$$+ m(I_q \otimes (vec\Sigma \otimes I_{q^2})) \frac{\partial}{\partial t^T} vec(\Sigma \otimes \Sigma t) + m(I_q \otimes (vec\Sigma \otimes I_{q^2})) \frac{\partial}{\partial t^T} vec(\Sigma t \otimes \Sigma)$$
$$= mvec\{(vec\Sigma \otimes I_{q^2})[(\Sigma \otimes \Sigma t) + (\Sigma t \otimes \Sigma)]\} \otimes (t^T\Sigma)$$
$$+ m(I_q \otimes (vec\Sigma \otimes I_{q^2}))[(vec\Sigma \otimes \Sigma) + (K_{q,q} \otimes I_q)(\Sigma \otimes vec\Sigma)]$$

The sixth moment of a multivariate normal distribution is, by setting t of the sixth derivative to 0,

$$E(t^6) = (\Sigma \otimes vec(\Sigma \otimes vec(\Sigma))) + (vec\Sigma \otimes I_{q^3})(\Sigma \otimes vec(\Sigma))$$
$$+ [(\Sigma \otimes (vec\Sigma)^T \otimes I_{q^4}](K_{q^3,q} \otimes I_{q^3})(\Sigma \otimes vec(I_{q^3})) + (vec\Sigma \otimes I_{q^3})(vec\Sigma \otimes \Sigma)$$
$$+ ((vec\Sigma)^T \otimes \Sigma \otimes I_{q^4})(K_{q^3,q} \otimes I_{q^3})(\Sigma \otimes vec(I_{q^3})) + (\Sigma \otimes vec[vec(\Sigma) \otimes \Sigma])$$
$$+ (vec\Sigma \otimes I_{q^3})(K_{q,q} \otimes I_q)(\Sigma \otimes vec\Sigma)$$
$$+ [(\Sigma \otimes (vec\Sigma)^T)(K_{q,q} \otimes I_q) \otimes I_{q^4}](K_{q^3,q} \otimes I_{q^3})(\Sigma \otimes vec(I_{q^3}))$$
$$+ [\Sigma \otimes vec\{(K_{q,q} \otimes I_q)(\Sigma \otimes vec(\Sigma))\}]$$
$$+ (\Sigma \otimes (vec\Sigma)^T)(K_{q,q} \otimes I_q) \otimes I_{q^4})(I_q \otimes K_{q^2,q} \otimes I_{q^3})(vec(I_q) \otimes I_{q^5})(K_{q^2,q} \otimes I_{q^2}) \times$$
$$(\Sigma \otimes vec(I_{q^2})) + [(\Sigma \otimes (vec(I_{q^2}))^T)(K_{q,q^2} \otimes I_{q^2}) \otimes I_{q^4}][\Sigma \otimes vec(\Sigma \otimes I_{q^3})]$$

$$+((vec\Sigma)^T \otimes \Sigma \otimes I_{q^4})(I_q \otimes K_{q^2,q^2} \otimes I_{q^3})(vec(I_q) \otimes I_{q^5})(K_{q^2,q^2} \otimes I_{q^2})(\Sigma \otimes vec(I_{q^2}))$$

$$+m\{[(\Sigma \otimes (vec(I_{q^2}))^T)(K_{q,q^2} \otimes I_{q^2})] \otimes I_{q^4}\}(I_{q^2} \otimes K_{q^3,q} \otimes I_{q^3})(I_{q^3} \otimes vec(I_{q^3}))(vec\Sigma \otimes \Sigma)$$

$$+(I_q \otimes (vec\Sigma \otimes I_{q^2}))[(vec\Sigma \otimes \Sigma)+(K_{q,q} \otimes I_q)(\Sigma \otimes vec\Sigma)].$$

APPENDIX C

APPENDIX C

Proof of the Substitutions

This Appendix will prove the substitutions of the expectations of the fourth and sixth moments with simpler forms.

The Fourth Moment Substitution

We have $[(b_i-\tilde{b}_i)^T\otimes(b_i-\tilde{b}_i)^T\otimes(b_i-\tilde{b}_i)^T]\tilde{l}_i^{(4)}(b_i-\tilde{b}_i)$

in the approximate marginal likelihood, where the $q\times 1$ vector $(b_i-\tilde{b}_i)$ has a multivariate normal distribution, $N(0,\Sigma)$, $\Sigma=(D^{-1}-\tilde{l}_i^{(2)})^{-1}$. For simplicity, I will use Σ for the derivation, use b for $(b_i-\tilde{b}_i)$ and ignore the subscript in $\tilde{l}_i^{(4)}$ here:

$$r=\left[b^T\otimes b^T\otimes b^T\right]\tilde{l}^{(4)}b. \tag{1}$$

Theorem 1

$$E(r)=3[vec(\Sigma\otimes\Sigma)]^T vec(\tilde{l}^{(4)}). \tag{2}$$

Proof The dimension of the first part in r, $\left[b^T\otimes b^T\otimes b\right]\tilde{l}^{(4)}$, is $1\times q$. The second part, b, is $q\times 1$. Since r is a scalar, by using PA-7 and regarding the first part as a^T and the second part as b, it becomes $r=tr\left\{b\left[b^T\otimes b^T\otimes b^T\right]\tilde{l}^{(4)}\right\}$. (3)

Separate the inside of the trace function into two parts, $r_1=b\left[b^T\otimes b^T\otimes b^T\right]$, a $q\times q^3$ matrix, and $r_2=\tilde{l}_i^{(4)}$, a $q^3\times q$ matrix. Using Equation PA-8, (3) becomes

$$r=\left(vec\left\langle(b\otimes b\otimes b)b^T\right\rangle\right)^T vec(\tilde{l}^{(4)})=(vec(r_1^T))^T vec(r_2). \tag{4}$$

Because $\tilde{l}^{(4)}$ is a constant, and b has a multivariate normal distribution, $N(0,\Sigma)$, the expectation of (4) becomes

$$E(r) = E\left((vec[(b\otimes b\otimes b)b^T])^T\right)vec(\tilde{l}^{(4)}) = E((vec(r_1^T))^T)vec(r_2), \quad (5)$$

where $E(r_1^T)$ is the fourth moment of a normal distribution, m_4. Therefore, the expectation of (1) is $E(r) = [vec(m_4)]^T vec(\tilde{l}^{(4)})$, (6)

$$m_4 = (\Sigma\otimes vec(\Sigma)) + [vec(\Sigma)\otimes\Sigma] + \{(K_{qq}\otimes I_q)(\Sigma\otimes vec(\Sigma))\}$$

$$= E_1 + E_2 + E_3 \quad (7)$$

where K_{qq} is a commutation matrix that permutes rows of a matrix, and I is an identity matrix. (See APPENDIX B: The Six Moments of Multivariate Normal Distribution)

Each matrix E_1, E_2, and E_3, in m_4 is a Kronecker product of $vec\Sigma$ and Σ. By the definition of the Kronecker product that $A\otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \ldots a_{1p}B \\ a_{21}B & a_{22}B & \ldots a_{2p}B \\ & \ldots\ldots & \\ a_{q1}B & a_{q2}B & \ldots a_{qp}B \end{bmatrix}$, the sum of the elements inside E_1 is equal to that inside E_2 as well as that inside E_3, which, in turn, are all the same as that inside $\Sigma\otimes\Sigma$. Therefore, m_4 has exactly the same sum of elements as $3(\Sigma\otimes\Sigma)$. Moreover, according to Anderson (1958), the expectation of the fourth moment of a multivariate normal is

$$E(b_k b_l b_m b_r) = \sigma_{kl}\sigma_{mr} + \sigma_{km}\sigma_{lr} + \sigma_{kr}\sigma_{ml}, \quad (8)$$

where σ_{kl} is the covariance of b_k and b_l; k, l, m, r can be equal. Note that the power of $b_k b_l b_m b_r$ corresponds to that of any of the $\sigma_{kl}\sigma_{mr}$, $\sigma_{km}\sigma_{lr}$, and $\sigma_{kr}\sigma_{ml}$, even though they are three different products of covariance terms.

On the other hand, since (1) is the fourth order term (without the factorial) of the fourth order Taylor expansion, with $\tilde{l}^{(4)}$ the fourth derivative of the likelihood, when multiplied out, it becomes:

$$E(r) = E\left(\sum_{k=1}^{q} b_k^4 \sum_{j=1}^{n_j} g_j z_{jk}^4 + 4\sum_{k\neq l}^{q} b_k^3 b_l \sum_{j=1}^{n_j} g_j z_{jk}^3 z_{jl} + 6\sum_{k<l}^{q} b_k^2 b_l^2 \sum_{j=1}^{n_j} g_j z_{jk}^2 z_{jl}^2 \right.$$
$$\left. + 12 \sum_{k\neq l\neq m, l<m}^{q} b_k^2 b_l b_m \sum_{j=1}^{n_j} g_j z_{jk}^2 z_{jl} z_{jm} + \sum_{k\neq l\neq m\neq r}^{q} b_k b_l b_m b_r \sum_{j=1}^{n_j} g_j z_{jk} z_{jl} z_{jm} z_{jr} \right)$$

$$= 3\sum_{k=1}^{q} \sigma_{kk}^2 \sum_{j=1}^{n_j} g_j z_{jk}^4 + 12\sum_{k\neq l}^{q} \sigma_{kk}\sigma_{kl} \sum_{j=1}^{n_j} g_j z_{jk}^3 z_{jl} + 6\sum_{k<l}^{q} (\sigma_{kk}\sigma_{ll} + 2\sigma_{kl}^2) \sum_{j=1}^{n_j} g_j z_{jk}^2 z_{jl}^2$$
$$+12 \sum_{k\neq l\neq m, l\neq m}^{q} (\sigma_{kk}\sigma_{lm} + 2\sigma_{kl}\sigma_{km}) \sum_{j=1}^{n_j} g_j z_{jk}^2 z_{jl} z_{jm} \qquad (9)$$
$$+ \sum_{k\neq l\neq m\neq r}^{q} (\sigma_{kl}\sigma_{mr} + \sigma_{km}\sigma_{lr} + \sigma_{kr}\sigma_{lm}) \sum_{j=1}^{n_j} g_j z_{jk} z_{jl} z_{jm} z_{jr}$$

Note that the power of z_{jk} also corresponds to the power of b_k, and the power of z_{jl} corresponds to the power of b_l, and so on. That is, $\sum_{k\neq l\neq m\neq r} g_j z_{jk}^{s_1} z_{jl}^{s_2} z_{jm}^{s_3} z_{jr}^{s_4}$, $s_1, s_2, s_3, s_4 = 0,..,4$, $s_1 + s_2 + s_3 + s_4 = 4$, is multiplied by $b_k^{s_1} b_l^{s_2} b_m^{s_3} b_r^{s_4}$. Thus the power of $z_k^{s_1} z_l^{s_2} z_m^{s_3} z_r^{s_4}$'s corresponds to that of $b_k^{s_1} b_l^{s_2} b_m^{s_3} b_r^{s_4}$'s, which in turn corresponds to that of the products of the variances and/or covariances as we see above. We have known that the sum of the terms in $3(\Sigma \otimes \Sigma)$ is equal to that of m_4. Therefore, to prove the substitution,

we will have to prove that the power of the product of covariances and/or variances in each position of $3[vec(\Sigma \otimes \Sigma)]^T$ corresponds to that of $z_k^{s_1} z_l^{s_2} z_m^{s_3} z_r^{s_4}$ in $\tilde{l}_i^{(4)}$, so that $3[vec(\Sigma \otimes \Sigma)]^T vec\tilde{l}^{(4)}$ will also result in (9).

Because $(vec\ m_4)^T$ is the expectation of $\left(vec\left((b \otimes b \otimes b)b^T\right)\right)^T$, the power of the sum of the three products of variances and/or covariances in each position of the former ($(vec\ m_4)^T$) correspond to the power in each position of the $b_s b_t b_u b_v$'s in the latter ($\left(vec\left((b \otimes b \otimes b)b^T\right)\right)^T$). On the other hand, because $\Sigma = E(bb^T)$, the power of the products of variances and/or covariances in each position of $3(\Sigma \otimes \Sigma)$ will correspond to that in each position of $(bb^T \otimes bb^T)$. Combining these two facts, the proof reduces to proving that $[vec(r_1^T)]^T = \left[vec\left((b \otimes b \otimes b)b^T\right)\right]^T = [vec(bb^T \otimes bb^T)]^T$.

Since to transpose a vector will not change the order of the elements in the vector, I will ignore the transpose on both sides of the above equation and prove:

$$vec\left((b \otimes b \otimes b)b^T\right) = vec(r_1^T) = vec(bb^T \otimes bb^T). \qquad (10)$$

By Equation PA-3, $\quad r_1^T = [b \otimes b \otimes b] \otimes b^T. \qquad (11)$

Putting back (11) to the left hand side of (10) and by Equations PA-11 and PA-6,

$$\mathrm{vec}(r_1^T) = vec[b \otimes b \otimes b \otimes b^T] = vec[b \otimes c \otimes b^T] = vec[b \otimes (c \otimes b^T)] = vec[b \otimes (b^T \otimes c)]$$

$$= vec[b \otimes b^T \otimes c] = vec[(b \otimes b^T) \otimes c] = \mathrm{vec}(\mathrm{b} \otimes \mathrm{b}^{\mathrm{T}}) \otimes \mathrm{c}. \qquad (12)$$

where $c = (b \otimes b)$, c being a $q^2 \times 1$ vector. The last equation of the third line in (12) is obtained by PA-12 by regarding $(\mathrm{b} \otimes \mathrm{b}^{\mathrm{T}})$ as a matrix.

Similarly, the right hand side of (10) is

$$vec[bb^T \otimes bb^T] = vec[(b \otimes b^T) \otimes (b \otimes b^T)] = vec[(b^T \otimes b) \otimes (b^T \otimes b)]$$
$$= vec[b^T \otimes (b \otimes b^T) \otimes b] = vec[b^T \otimes (b^T \otimes b) \otimes b] = vec[b^T \otimes b^T \otimes b \otimes b] \qquad (13)$$
$$= vec[(b^T \otimes b^T) \otimes c] = vec(b^T \otimes b^T) \otimes c = vec(b \otimes b^T) \otimes c,$$

which is equal to the left hand side.

Therefore I have proved Equation 10, and thus the substitution.

The Sixth Moment Substitution

Theorem 2 $$Etr\left\{(b^T \otimes b^T \otimes b^T \otimes b^T \otimes b^T)\tilde{l}_i^{(6)} b\right\}$$
$$= 15[vec(\Sigma \otimes \Sigma \otimes \Sigma)]^T vec\tilde{c}\, l_i^{(6)}$$

The theorem can be proved by following exactly the same reasoning as above.

Theorem 3 $$Etr\left\{(b \otimes b)\left[\left[vec(bb^T)\right]^T \otimes \left[vec(bb^T)\right]^T\right]\left(\tilde{l}_i^{(3)} \otimes \tilde{l}_i^{(3)}\right)\right\}$$
$$= 15tr\left\{vec\Sigma\left(\left[vec\Sigma\right]^T \otimes \left[vec\Sigma\right]^T\right)\left(\tilde{l}_i^{(3)} \otimes \tilde{l}_i^{(3)}\right)\right\} \qquad (15)$$

Proof The argument for the substitution is similar to the above, too. That is, each of the fifteen different matrices in the sixth moment (See APPENDIX B: The Six Moments of a Multivariate Normal Distribution) is the Kronecker product of the variance matrix arranged in a special way by using identity matrices and commutation matrices. Therefore, the total of the elements inside each of the matrices will be the same as that of.

$vec\Sigma\left(\left[vec\Sigma\right]^T \otimes \left[vec\Sigma\right]^T\right)$. Since Equation 15 is a scalar, the usefulness of the identity matrices and commutation matrices in the sixth moment disappears. I only need to show that

$$vec\left\langle (b\otimes b)\left[\left[vec(bb^T)\right]^T \otimes \left[vec(bb^T)\right]^T\right]\right\rangle = vec\left(b\otimes b\otimes b\otimes b\otimes b\otimes b^T\right). \quad (16)$$

The left hand side of Equation 16 can be simplified to

$$vec\left\langle (b\otimes b)\left[\left[vec(bb^T)\right]^T \otimes \left[vec(bb^T)\right]^T\right]\right\rangle$$
$$= vec\left\langle (b\otimes b)\left[b^T\otimes b^T\otimes b^T\otimes b^T\right]\right\rangle = b\otimes b\otimes b\otimes b\otimes b\otimes b,$$

which is also the simplified version of the right hand side. Hence we can substitute the fifteen matrices of the sixth moment with $15vec\Sigma\left(\left[vec\Sigma\right]^T \otimes \left[vec\Sigma\right]^T\right)$.

APPENDIX D

APPENDIX D

The Expectation of the Third Order Term Squared

This appendix will find the expectation of the third order term squared in the approximation to the Taylor series inside the exponential. To simplify equations, I will use the same notation as in APPENDIX C. That is, $b = b_i - \tilde{b}_i$, and $b \sim N(0, \Sigma)$, with $\Sigma = (D^{-1} - \tilde{l}_i^{(2)})^{-1}$. Thus, this appendix will show that the expectation of L_3^2 is

$$E(L_3^2) = \frac{15}{72}\left(\sum_j^{n_i} a_{ij} Z_{ij}^T \Sigma Z_{ij} Z_{ij}^T\right)\Sigma\left(\sum_j^{n_i} a_{ij} Z_{ij} Z_{ij}^T \Sigma Z_{ij}\right), \text{ where } L_3^2 = \frac{1}{2}\left(\frac{1}{3!}\left[b^T \otimes b^T\right]\tilde{l}_i^{(3)} b\right)^2,$$

with $\tilde{l}_i^{(3)} = -(Z_i^T \otimes Z_i^T) A_i Z_i$.

Proof

Since L_3^2 is a scalar, a trace function of the scalar will not change the value.

$$L_3^2 = \frac{1}{72}\left[tr\left(b\left[(b^T \otimes b^T\right]\tilde{l}_i^{(3)}\right)\right]^2, \text{ then by PA-13,}$$

$$= \frac{1}{72} tr\left[\left(\left(b\left[b \otimes b^T\right]\tilde{l}_i^{(3)}\right) \otimes \left(b\left[b^T \otimes b^T\right]\tilde{l}_i^{(3)}\right)\right]\right.$$

$$= \frac{1}{72} tr\left[\left\langle\left(b\left[b^T \otimes b^T\right]\right) \otimes \left(b\left[b^T \otimes b^T\right]\right)\right\rangle \left(\tilde{l}_i^{(3)} \otimes \tilde{l}_i^{(3)}\right)\right] \text{ (by PA-14)}$$

$$= \frac{1}{72} tr\left[\left\langle(b \otimes b)\left[\left[b^T \otimes b^T\right] \otimes \left[b^T \otimes b^T\right]\right]\right\rangle \left(\tilde{l}_i^{(3)} \otimes \tilde{l}_i^{(3)}\right)\right].$$

Take expectation of the above function and substitute the sixth moment with $15 vec\Sigma\left(\left[vec\Sigma\right]^T \otimes \left[vec\Sigma\right]^T\right)$. (See APPENDICES B and C).

$$E(L_3^2) = \frac{15}{72} tr\left[\left\langle vec\Sigma\left(\left[vec\Sigma\right]^T \otimes \left[vec\Sigma\right]^T\right)\right\rangle \left(\tilde{l}_i^{(3)} \otimes \tilde{l}_i^{(3)}\right)\right]$$

$$= \frac{15}{72} tr\left\langle vec\Sigma\left(\left[vec\Sigma \right]^T \tilde{l}_i^{(3)} \otimes \left[vec\Sigma \right]^T \tilde{l}_i^{(3)} \right) \right\rangle \quad \text{(by PA-14)}$$

$$= \frac{15}{72} \left(\left\langle \left[vec\Sigma \right]^T \tilde{l}_i^{(3)} \right\rangle \otimes \left\langle \left[vec\Sigma \right]^T \tilde{l}_i^{(3)} \right\rangle \right) vec\Sigma \quad \text{(by PA-15 and PA-8)}$$

$$= \frac{15}{72} \left[vec\Sigma \right]^T \tilde{l}_i^{(3)} \Sigma \tilde{l}_i^{(3)T} vec\Sigma. \qquad \text{(by PA-16)}$$

Furthermore, $\tilde{l}_i^{(3)} = -(Z_i^T \otimes Z_i^T) A_i Z_i$ and $A_i = \sum_j^{n_i} a_{ij} (E_j E_j^T \otimes E_j)$, where E_j is an $n_i \times 1$ vector with the jth element being 1 and the others being 0. Thus,

$$\tilde{l}_i^{(3)} = -(Z_i^T \otimes Z_i^T) \left[\sum_j^{n_i} a_{ij} (E_j E_j^T \otimes E_j) \right] Z_i = \sum_j^{n_i} a_{ij} (Z_{ij} E_j^T \otimes Z_{ij}) (Z_i \otimes 1)$$

$$= \sum_j^{n_i} a_{ij} (Z_{ij} Z_{ij}^T \otimes Z_{ij}).$$

Therefore, $E(L_3^2)$ becomes, by applying PA-16 and its transpose to

$$\left[\sum_j^{n_i} a_{ij} (Z_{ij} Z_{ij}^T \otimes Z_{ij}) \right]^T vec\Sigma \text{ and } \left[vec\Sigma \right]^T \left[\sum_j^{n_i} a_{ij} (Z_{ij} Z_{ij}^T \otimes Z_{ij}) \right] \text{ respectively,}$$

$$E(L_3^2) = \frac{15}{72} \left[vec\Sigma \right]^T \left[\sum_j^{n_i} a_{ij} (Z_{ij} Z_{ij}^T \otimes Z_{ij}) \right] \Sigma \left[\sum_j^{n_i} a_{ij} (Z_{ij} Z_{ij}^T \otimes Z_{ij}) \right]^T vec\Sigma$$

$$= \frac{15}{72} \left(\sum_j^{n_i} a_{ij} Z_{ij}^T \Sigma Z_{ij} Z_{ij}^T \right) \Sigma \left(\sum_j^{n_i} a_{ij} Z_{ij} Z_{ij}^T \Sigma Z_{ij} \right).$$

APPENDIX E

APPENDIX E

Computational Algorithm

To adapt to the PQL computer program (Bryk, Raudenbush and Congdon, 1996), the notations for equations in Chapter 2 are changed as follows:

Chapter 2	D	X_i	Z_i	β	b_i
APPENDIX E	T	A_{1j}	A_{2j}	θ_1	θ_2

The approximate log-likelihood is

$$\sum_{j=1}^{J}\left\{\frac{-1}{2}\log|\mathrm{T}|-\frac{1}{2}\log|\mathrm{C_j}|+\tilde{\mathrm{l}}_{\mathrm{j}}-\frac{1}{2}\theta_{2\mathrm{j}}^{\mathrm{T}}\mathrm{T}^{-1}\theta_{2\mathrm{j}}+\log \mathrm{D_j}\right\} \qquad (1)$$

where $C_j = \mathrm{T}^{-1} + A_{2j}^T W_j A_{2j}$;

$\tilde{l}_j = y_j^T \eta_j + \sum_i \log(1-\mu_{ij})$, y_j is an $n_j \times 1$ column vector of the response of the jth cluster;

η_j is an $n_j \times 1$ column vector for cluster j with the ith component being η_{ij},

$\eta_{ij} = A_{1ij}^T \theta_1 + A_{2ij}^T \theta_{2j}$, and $\mu_{ij} = \dfrac{1}{1+\exp(-\eta_{ij})}$;

θ_1 is from previous iteration;

W_j is an $n_j \times n_j$ diagonal matrix with the ith diagonal term being $w_{ij} = \mu_{ij}(1-\mu_{ij})$;

$Z_j = W_j^{-1}(y_j - \mu_j) + \eta_j$ is the linearized dependent variable of group j;

μ_j is an $n_j \times 1$ column vector for cluster j with the ith component being μ_{ij};

A_{2j} is an $n_j \times q$ random effect design matrix for the jth group;

A_{2ij} is a $q \times 1$ column vector for observation i in group j, the transpose of the ith row of A_{2j};

$g_{ij} = w_{ij}(1-6w_{ij})$.

$a_{ij} = w_{ij}(1-2\mu_{ij})$, $i = 1, 2, ..., n_j$;

$B_{ij} = A_{2ij}^T C_j^{-1} A_{2ij}$;

A_{1ij} is a $p \times 1$ column vector for observation i in group j, the transpose of the ith row of A_{1j};

$h_{ij} = a_{ij}(1-12w_{ij})$;

$r_{ij} = g_{ij}(1-12w_{ij}) - 12a_{ij}^2$;

$$k_j = \sum_i^{n_j} a_{ij} A_{2ij} B_{ij};$$

$$D_j = 1 - \frac{1}{8}\sum_i^{n_j} g_{ij} B_{ij}^2 - \frac{1}{48}\sum_i^{n_j} r_{ij} B_{ij}^3 + \frac{15}{72} k_j^T C_j^{-1} k_j.$$

The **score function for** β **in the log-likelihood is**

$$S_{\theta_1 j} = \frac{-1}{2}\sum_i a_{ij} B_{ij} A_{1ij} + \frac{1}{2}\sum_i a_{ij} B_{ij} A_{1j}^T W_j A_{2j} C_j^{-1} A_{2ij}$$

$$+ A_{1j}^T W_j Z_j - A_{1j}^T W_j A_{2j} \theta_{2j} - A_{1j}^T W_j A_{1j} \theta_1$$

$$+ \frac{1}{D_j}\left\langle -\frac{1}{8}\sum_i h_{ij} B_{ij}^2 A_{1ij} - \frac{1}{48}\sum_i^{n_j} p_{ij} B_{ij}^3 A_{1ij} + \frac{15}{36}\sum_i^{n_j} g_{ij} B_{ij} s_{ij} A_{1ij} - \frac{15}{72}\sum_i^{n_j} a_{ij} s_{ij}^2 A_{1ij} \right.$$

$$+ \frac{1}{4}\sum_i \sum_h a_{hj} B_{ij} g_{ij} (A_{2hj}^T C_j^{-1} A_{2ij})^2 A_{1hj} + \frac{1}{16}\sum_h^{n_j}\sum_i^{n_j} a_{hj} r_{ij} B_{ij}^2 (A_{2hj}^T C_j^{-1} A_{2ij})^2 A_{1hj}$$

$$-\frac{15}{36}\sum_{i}^{n_j}\sum_{h}^{n_j} a_{hj}a_{ij}s_{ij}((A_{2hj}^{T}C_j^{-1}A_{2ij})^2 A_{1hj})$$

$$-\frac{1}{4}\sum_{i}\sum_{h} a_{hj}B_{ij}g_{ij}(A_{2hj}^{T}C_j^{-1}A_{2ij})^2(A_{1j}^{T}W_jA_{2j}C_j^{-1}A_{2hj})$$

$$+\frac{1}{48}\sum_{i}^{n_j} p_{ij}B_{ij}^3A_{1j}^{T}W_jA_{2j}C_j^{-1}A_{2ij}$$

$$-\frac{1}{16}\sum_{h}^{n_j}\sum_{i}^{n_j} a_{hj}r_{ij}B_{ij}^2(A_{2hj}^{T}C_j^{-1}A_{2ij})^2(A_{1j}^{T}W_jA_{2j}C_j^{-1}A_{2hj})$$

$$-\frac{15}{36}\sum_{i}^{n_j} g_{ij}B_{ij}s_{ij}(A_{1j}^{T}W_jA_{2j}C_j^{-1}A_{2ij})+\frac{1}{8}\sum_{i} h_{ij}B_{ij}^2A_{1j}^{T}W_jA_{2j}C_j^{-1}A_{2ij}$$

$$+\frac{15}{36}\sum_{i}^{n_j}\sum_{h}^{n_j} a_{hj}a_{ij}s_{ij}((A_{2hj}^{T}C_j^{-1}A_{2ij})^2 A_{1j}^{T}W_jA_{2j}C_j^{-1}A_{2hj})$$

$$+\frac{15}{72}\sum_{i}^{n_j} a_{ij}s_{ij}^2A_{1j}^{T}W_jA_{2j}C_j^{-1}A_{2ij}\Bigg\rangle$$

$$=(A1PY)_j-(A1PA1\theta_1)_j-(A1PA2)_j\theta_{2j}+\sum_{i}^{n_j} f_{ij}l_{ij}+\frac{1}{D_j}\sum_{i}^{n_j} c_{ij}l_{ij} \qquad (2)$$

where $(A1PY)_j = A_{1j}^{T}W_jZ_j$,

$(A1PA1)_j = A_{1j}^{T}W_jA_{1j}$,

$(A1PA2)_j = A_{1j}^{T}W_jA_{2j}$,

$\mathrm{f_{ij}} = \frac{-1}{2}\mathrm{a_{ij}B_{ij}}$,

$l_{ij} = A_{1ij} - X_jA_{2ij}$,

$X_j = A_{1j}^{T}W_jA_{2j}C_j^{-1}$,

$$c_{ij} = -\frac{1}{8}h_{ij}B_{ij}^2 + \frac{1}{4}a_{ij}A_{2ij}^T m_j A_{2ij} - \frac{1}{48}p_{ij}B_{ij}^3 + \frac{1}{16}a_{ij}A_{2ij}^T b_j A_{2ij} + \frac{15}{36}g_{ij}B_{ij}s_{ij} - \frac{15}{72}a_{ij}s_{ij}^2$$

$$-\frac{15}{36}a_{ij}A_{2ij}^T e_j A_{2ij},$$

where $m_j = \sum_i^{n_j} B_{ij}g_{ij}C_j^{-1}A_{2ij}A_{2ij}^T C_j^{-1}$,

$p_{ij} = h_{ij}(1-12w_{ij}) - 36a_{ij}g_{ij}$,

$\mathbf{b_j} = \sum_{\mathbf{i}}^{\mathbf{n_j}} \mathbf{r_{ij}B_{ij}^2C_j^{-1}A_{2ij}A_{2ij}^T C_j^{-1}}$.

$\mathbf{s_{ij} = A_{2ij}^T C_j^{-1} k_j}$

$e_j = \sum_i^{n_j} a_{ij}s_{ij}C_j^{-1}A_{2ij}A_{2ij}^T C_j^{-1}$.

For variance components:

The **score function for** ϕ in the log-likelihood is

$$S_{\phi} = \frac{1}{2}E^T vec[\mathrm{T}^{-1}(\hat{\mathrm{T}}_j - \mathrm{T})\mathrm{T}^{-1}] \; -\frac{1}{2}\sum_i^{n_j} a_{ij}B_{ij}E^T vec[Q_{ij}]$$

$$+\frac{1}{D_j}\left\{\sum_i^{n_j} c_{ij}\; E^T vec[Q_{ij}]\right.$$

$$\left.+E^T\left[\sum_i^{n_j}(-\frac{1}{4}g_{ij}B_{ij} - \frac{1}{16}r_{ij}B_{ij}^2 + \frac{15}{36}a_{ij}s_{ij})vec(F_{ij}) + \frac{15}{72}vec\left\{\mathrm{T}^{-1}C_j^{-1}k_j k_j^T C_j^{-1}\mathrm{T}^{-1}\right\}\right]\right\} \quad (3)$$

where $\hat{\mathrm{T}}_j = \theta_{2j}\theta_{2j}^T + C^{-1}{}_j$, T is from previous iteration.

$E = \frac{dvec\mathrm{T}}{d\phi^T}$, ϕ being the unique elements in $vec\mathrm{T}$;

$(A2PR)_j = A_{2j}^T W_j Z_j - A_{2j}^T W_j A_{1j}\theta_1 - A_{2j}^T W_j A_{2j}\theta_{2j}$

$Q_{ij} = \mathrm{T}^{-1} C_j^{-1} A_{2ij} (A2PR)_j^T$

$F_{ij} = \mathrm{T}^{-1} C_j^{-1} A_{2ij} A_{2ij}^T C_j^{-1} \mathrm{T}^{-1}$.

The Fisher Scoring for the parameters is

$$\begin{bmatrix} \theta^{i+1} - \theta^i \\ vec\phi^{i+1} - vec\phi^i \end{bmatrix} = H^{-1} S \ , \quad S_j = \begin{bmatrix} S_{\theta j} \\ S_{\phi j} \end{bmatrix}, \ S = \sum_{j=1}^{J} S_j \quad H = \sum_{j=1}^{J} S_j S_j^T \ . \tag{4}$$

Program Flow:

1. Start with HLM2 estimates of T and θ_1.

2. Iteratively solve $C_j^{-1}(A_{2j}^T W_j Z_j - A_{2j}^T W_j A_{1j} \theta_i) = \theta_{2j}$ for θ_{2j} with W_j and z_j computed holding constant T and θ_1 from the last iteration. Thus, Z_j, η_j, μ_j, w_j, $(A2PA2)_j$, $(A2PY)_j$, C_j^{-1}, and $(A2PA1\theta_1)_j$ along with θ_{2j} will result from this iterative process.

3. Compute $(A1PA1)_j$, $(A1PA2)_j$, and $(A1PY)_j$ from w_j.

4. Compute a_{ij}, g_{ij}, h_{ij}, r_{ij} and p_{ij} from w_j and μ_j.

5. Compute B_{ij} from C_j^{-1}.

6. Compute f_{ij} and k_j from a_{ij} and B_{ij}.

7. Compute m_j from B_{ij} and C_j^{-1}.

8. Compute s_{ij} from k_j and C_j^{-1}.

9. Compute b_j from r_{ij}, B_{ij} and C_j^{-1}.

10. Compute e_j from C_j^{-1}, s_{ij}, and a_{ij}.

11. Compute c_{ij} from B_{ij}, m_j, b_j, a_{ij}, e_j, h_{ij}, and p_{ij}.

12. Compute $(A2PR)_j$ from $(A2PY)_j$, $(\text{A2PA1}\theta_1)_j$, and $(\text{A2PA2})_j$.

13. Compute X_j and l_{ij} from $(A1PA2)_j$ and C_j^{-1}.

14. Compute $\hat{T}$ from θ_2 and C_j^{-1}.

15. Compute Q_{ij} from C_j^{-1} and $(A2PR)_j$.

16. Compute F_{ij} from C_j^{-1}.

17. Complete the 2 S's and hence the new T, θ_1.

18. Monitor convergence as in current HLM. At convergence compute standard errors from square roots of diagonal elements of H^{-1}.

LIST OF REFERENCES

LIST OF REFERENCES

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

Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis.* New York: John Wiley & Sons, Inc..

Bennett, 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 generalised 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, T. (1996). *HLM2 and HLM3: Computer Programs and User's Guide*, v. 4.25, Chicago: Scientific Software International.

Chan, K. S. and Ledholter, J. (1995). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9-25.

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.

Fellner, W. H. (1986). Robust estimation of variance components. *Technometrics* 28, 51-60.

Fellner, W. H. (1987). Sparse matrices, and the estimation of variance components by likelihood methods. *Communication in Statistics*, B 16, 439-63.

Fulks, W. (1978). *Advanced Calculus: an Introduction to Analysis* (3rd ed.). New York: John Wiley & Sons.

Gelfand, A. E., and Charlin, B. P. (1993). Maximum likelihood estimation for constrained- or missing-data problems. *Canadian Journal of Statistics*, 21, 303-11.

Gelfand, A. E., Hills, S., Racine-Poon, A., and Smith, A. F. M. (1990). Illustartion of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, 85, 972-85.

Geyer, C. J., and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society* B, 54, 657-99.

Gelfand, A. E., and Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association,* 85, 398-409.

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine intelligence,* 6, 721-41.

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. and Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society,* A, 159, 505-13.

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives, *Journal of the Royal Society*, B, 46,149-192.

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 Programs in Biomedicine*, 49, 157-176.

Hobert, J., and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, 91, 1461-73.

Laird, N. M., and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrika*, 38, 963-74.

Lee, V. E. (1995). Another look at high school restructuring. More evidence that it improves student achievement and more insight into why. *Issues in Restructuring Schools*, 9,1-10.

Lee, Y., and Nelder, J. A. (1996). Hierarchical Generalized Linear Models. *Journal of the Royal Statistical Society*, B, 58, 619-78.

Lin, X. and Breslow, N. E. (1996), "Bias Correction in Generalized Linear Mixed Models with Multiple Components of Dispersion," *Journal of the American Statistical Association*, 91, 1007-1016.

Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, 74, 812-27.

Longford, N. T. (1988a). A quasi-likelihood adaptation for variance component analysis. *Proceedings of the Statistical Computing, American Statistical Association.*

Longford, N. T. (1988b). *VARCL: Software for Variance Component Analysis of Data with Hierarchically Nested Random Effects (Maximum Likelihood.)* Princeton: Educational Testing Service.

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

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

Mac Iver, D. J., and Plank, S. B. (1996). The Talent Development Middle School. Creating a motivational climate conducive to talent development in middle schools: implementation and effects of Student Team Reading. (Report No. 4). Center for Research on the Education of Studnts Placed at Risk. (ERIC Document Reproduction Service No. ED 402 388)

Magnus, J. R. (1988). *Linear Structures*, New York: Oxford University Press.

Magnus, J. R., and Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*, New York: John Wiley & Sons.

Marks, H. M. (1995). Student engagement in the classrooms of restructuring schools. Center on Organization and Restructuring of Schools. (ERIC Document Reproduction Service No. ED 381 884)

Mason, W. M., Wong, G. M., and Entwistle, B. (1983). Contextual analysis through the multilevel linear model. In *Sociological Methodology*, 1983-1984, 72-103.

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, 162-70.

Natarajan, R., and McCulloch, C. E. (1995). A note on the existence of the posterior distribution for a class of mixed models for binomial responses. *Biometrika*, 82, 639-43.

Nelder, J. A. and Wedderburn, R. W. M. (1982). Generalized linear models. *Journal of the Royal Statistical Society*, A, 135, 370-84.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58, 545-54.

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, 12-35.

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

Raudenbush, S. W. (1988). Educational applications of hierarchical linear models: a review. *Journal of Educational Statistics*, 13, 85-116.

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

Raudenbush, S. W. and Bhumirat, C. and Kamali, M. (1992). Predictors and consequences of primary teachers' sense of efficacy and students' perceptions of achievement in Thailand. *International Journal of Educational Research*, 17, 165

-177.

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 Chan, W. (1993). Application of a hierarchical linear model to the study of adolescent deviance in an overlapping cohort design. *Journal of Consulting and Clinical Psychology*, 61, 941-51.

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

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

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

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78, 719-27.

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

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.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models. *Applied Statistics*, 30, 125-31.

Willms, J. D., and Raudenbush, S. W. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement*, 26, 209-32.

Yang, M. (1994). A simulation study for the assessment of the non-linear hierarchical model estimation via approximate maximum likelihood. Unpublished apprenticeship paper: 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 aaper: Michigan State University.

Young, D. J. (1996). Science achievement and educational productivity: a hierarchical linear model. *Journal of Educational Research*, 8, 272-78.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association*, 86, 79-86.

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