UNIVERSITY OF CALIFORNIA,
IRVINE

New Approaches to Model Selection in Bayesian Mixed Modeling

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Statistics

by

Fletcher G.W. Christensen

Dissertation Committee:
Professor Wesley O. Johnson, Chair
Professor Edward J. Bedrick
Associate Professor Michele Guindani

2017

DEDICATION

To my parents: For their boundless reserves of patience, humor, and helpfulness.

TABLE OF CONTENTS

Page

LIST OF FIGURES vi

LIST OF TABLES vii

ACKNOWLEDGMENTS viii

CURRICULUM VITAE ix

ABSTRACT OF THE DISSERTATION xi

1 Introduction 1

2 Environmental Effects on Human Fertility and Shared Parameter Modeling 5

2.1 Environmental Effects on Human Fertility 6
2.1.1 Methods 7
2.1.2 Results 19
2.1.3 Discussion 32
2.2 Shared Parameter Modeling 35
2.2.1 Background 35
2.2.2 Shared Parameter Modeling 40
2.2.3 Simulation Results 43
2.2.4 Future work 48

3 Marginalization for DIC – Part I 49

3.1 Background 49
3.1.1 Philosophy of Model Selection 50
3.1.2 Kullback-Leibler (KL) Divergence, 1951 51
3.1.3 Akaike's Information Criterion (AIC), 1974 52
3.1.4 Bayes Factors (BF) 55
3.1.5 Bayesian Information Criterion (BIC), 1978 56
3.1.6 Log Pseudo-Marginal Likelihood (LPML), 1979 59
3.1.7 Deviance Information Criterion (DIC), 2002 60
3.1.8 Other Information Criteria 65

3.2 Model Selection in Mixed Models . . . 67
3.2.1 The Mixed Modeling Framework . . . 67
3.2.2 Complications with p_D in Mixed Models . . . 71
3.3 The Need for Marginalization . . . 78
3.3.1 Three DIC Constructions for Mixed Models . . . 79
3.3.2 Why Do We Prefer the Marginal DIC? . . . 84
3.3.3 Arguments Against Marginalization . . . 90
3.4 Marginalization in the Linear Mixed Model . . . 93
3.4.1 Methods for Marginalization . . . 94

4 Marginalization for DIC – Part II **100**
4.1 Generalized Linear Mixed Models . . . 101
4.1.1 Development . . . 101
4.1.2 Use . . . 103
4.1.3 The Binomial and Poisson GLMMs . . . 104
4.2 Marginalization in GLMMs . . . 106
4.2.1 Preprocessing Won't Work . . . 106
4.2.2 Postprocessing Is Difficult . . . 107
4.2.3 Other Approaches Involving Hierarchical Models . . . 112
4.3 A Limited Marginalization Approach . . . 113
4.3.1 Approximate Marginalization through Taylor Expansion . . . 113
4.3.2 The Binomial and Poisson Approximations . . . 124
4.4 Simulation Results . . . 130
4.4.1 Description of Simulation Procedures . . . 130
4.4.2 Pseudo-KL Results . . . 133
4.4.3 Comparing the Approximation to Numerical Integration . . . 136
4.4.4 Simulated Stepwise Variable Selection . . . 143
4.5 Discussion . . . 146

5 Marginalization for DIC – Part III **147**
5.1 New Considerations . . . 148
5.1.1 REO vs. Non-REO GLMMs . . . 148
5.1.2 Why the Limited Approach Fails . . . 149
5.2 A General Marginalization Approach . . . 151
5.2.1 Approximate Marginalization through Root-Finding and Taylor Expansion . . . 152
5.2.2 The Binomial and Poisson Approximations . . . 158
5.3 Simulation Results . . . 161
5.3.1 Description of Simulation Procedures . . . 161
5.3.2 Pseudo-KL results . . . 166
5.3.3 Comparing the Approximation to Numerical Integration . . . 167
5.3.4 Simulated Stepwise Variable Selection . . . 172
5.4 Discussion . . . 172

6 Future Directions **175**
6.1 Future Research on Missing Data Methods 176
6.2 Future Research on GLMM Marginalization 177
6.2.1 Properties of the Marginal Approximation 177
6.2.2 Methods for Achieving Greater Computational Efficiency 178
6.2.3 Small-Sample Results When $q > 1$ 179
6.2.4 More Random Effects Distributions 183

A Appendix **193**
A.1 Complete the Square 193
A.2 Gauss-Hermite Quadrature 196
A.3 Taylor Approximation Error in REO GLMMs 199

LIST OF FIGURES

Page

2.1 Posterior density estimates for baseline coefficients 16
2.2 Representative LH and $E_1$3G concentrations for two participants 22
2.3 Graphical representation of changes in follicular phase length with changes in concentrations of pyrene and fluorene metabolites 29
2.4 Graphical representation of changes in highest LH with changes in concentrations of naphthalene and fluorene metabolites 31

3.1 An example of "skipping" in predictive error for LASSO LMM models 91

4.1 Percentage difference in DIC_m approximations 141

5.1 Example time-varying covariates 163

LIST OF TABLES

Page

2.1 Continuous characteristics of study participants 20
2.2 Discrete characteristics of study participants 21
2.3 Summary statistics for menstrual cycle endpoint variables 23
2.4 Summary statistics for OH-PAH concentrations 23
2.5 Correlations among concurrent measurements of OH-PAHs 25
2.6 Baseline and PAH models (1/2) 26
2.7 Baseline and PAH models (2/2) 27
2.8 Mean squared prediction errors: $k = 50$, $N^C = N^I = 3$, $\beta = \gamma = 1$ 46
2.9 Mean squared prediction errors: $\rho = 0.7$, $\phi = 0.0$, $\beta = \gamma = 1$ 46
2.10 Mean squared prediction errors: $\rho = 0.7$, $\phi = 0.0$, $N^C = N^I = 3$, $\gamma = 1$ 47

4.1 Pseudo-Kullback-Leibler divergences between true and approximated joint distributions 135
4.2 p_D's and DIC's for Bernoulli GLMM simulations 138
4.3 p_D's and DIC's for Poisson GLMM simulations 139
4.4 Runtimes for DIC approximations 142
4.5 Comparison of stepwise methods with different DIC approximations 145

5.1 Pseudo-Kullback-Leibler divergences between true and approximated joint distributions 166
5.2 p_D's and DIC's for binomial and Poisson GLMM simulations 169
5.3 Runtimes for DIC approximations 170
5.4 Comparison of stepwise methods with different DIC approximations 173

ACKNOWLEDGMENTS

I would like to thank the faculty, staff, and students of the statistics department at UC Irvine, to whom I am greatly indebted. In particular, I would like to thank Drs. Wesley O. Johnson, Michele Guindani, and Edward J. Bedrick, all of whom have been invaluable to my education in statistics at *many* stages.

Work presented in Chapter 2 of this dissertation was funded by the National Institutes of Health (NIH), National Institute of Environmental Health Sciences grant numbers ES016846 and ES020454 (principal investigator Dr. Ulrike Luderer); grant number UL1 RR031985 from the National Center for Research Resources(NCRR); a California Environmental Contaminant Biomonitoring Program RFI Laboratory Pilot Study; and the UC Irvine Center for Occupational and Environmental Health.

The text of this dissertation includes material reprinted from *Environment International*. Drs. Ulrike Luderer and Wesley Johnson directed and supervised this reprinted research, which was also assisted by Drs. Jianwen She, Ho Sai Simon Ip, Junqiang Zhou, Josephine Alvaran, Edward F. Krieg Jr, and James S. Kesner.

CURRICULUM VITAE

Fletcher G.W. Christensen

EDUCATION

Doctor of Philosophy in Statistics **2017**
University of California, Irvine *Irvine, CA*

Master of Arts in Statistics **2012**
University of New Mexico *Albuquerque, NM*

Postgraduate Diploma in Japanese Language and Culture **2006**
University of Sheffield *Sheffield, UK*

Bachelor of Arts in Psychology **2004**
University of Oklahoma *Norman, OK*

Bachelor of Science in Mathematics **2004**
University of Oklahoma *Norman, OK*

RESEARCH EXPERIENCE

Graduate Research Assistant **2014–2016**
University of California, Irvine *Irvine, CA*

Graduate Student Researcher **2010–2012**
Mind Research Network *Albuquerque, NM*

TEACHING EXPERIENCE

Teaching Associate **2016**
University of California, Irvine *Irvine, CA*

Teaching Assistant **2012–2017**
University of California, Irvine *Irvine, CA*

Teaching Assistant **2009–2010**
University of New Mexico *Albuquerque, NM*

Assistant Language Teacher **2004–2007**
Hida Takayama High School, Yamada Campus *Takayama, Japan*

REFEREED JOURNAL PUBLICATIONS

Ulrike Luderer et al. (2017). "Associations between urinary biomarkers of polycyclic aromatic hydrocarbon exposure and reproductive function during menstrual cycles in women". In: *Environment International* 100, pp. 110–120. DOI: 10.1016/j.envint.2016.12.021

Guoyi Zhang et al. (2015). "Nonparametric regression estimators in complex surveys". In: *Journal of Statistical Computation and Simulation* 85.5, pp. 1026–1034. DOI: 10.1080/00949655.2013.860139

Mustafa S. Çetin et al. (2014). "Thalamus and posterior temporal lobe show greater inter-network connectivity at rest and across sensory paradigms in schizophrenia". In: *NeuroImage* 97, pp. 117–126. DOI: 10.1016/j.neuroimage.2014.04.009

Ronald R. Christensen and Fletcher G.W. Christensen (2009). "Letters to the editor". In: *The American Statistician* 63.2, p. 197. DOI: 10.1198/tast.2009.0037

ABSTRACT OF THE DISSERTATION

New Approaches to Model Selection in Bayesian Mixed Modeling

By

Fletcher G.W. Christensen

Doctor of Philosophy in Statistics

University of California, Irvine, 2017

Professor Wesley O. Johnson, Chair

Because the marginal densities corresponding to data modeled with generalized linear mixed models (GLMMs) usually lack closed-form expressions, model selection via existing tools like the deviance information criterion (DIC) can yield inconsistent results. We discuss why marginalization is preferable for the evaluation of competing mixed models, provide a new method for fast and accurate approximation of the marginal DIC for GLMMs, and demonstrate through simulation how numerical approximation of the DIC relative to our marginalization scheme gives more accurate model selection results than other numerical approximation methods for DIC. We also discuss some issues related to model selection in an analysis of longitudinal data collected to assess the effect of polycyclic aromatic hydrocarbons on hormone functioning in women who were attempting to get pregnant.

Chapter 1

Introduction

This dissertation will develop methods to address two problems in Bayesian mixed modeling. The first method concerns how to make the best use of data where covariates of scientific interest are only available for half of response observations. The second method concerns how to approximate a marginalized form of the deviance information criterion (DIC; D. J. Spiegelhalter et al. 2002) for mixed models, particularly generalized linear mixed models with normally distributed random effects. Chapter 2 will explore the first question, along with providing a detailed, peer-reviewed analysis of environmental epidemiology data using the method we develop. Chapters 3, 4, and 5 will deal with the second question in progressively more general settings. Chapter 6 discusses future directions for our missing data work on the study discussed in Chapter 2, and for our methods developed to approximate a marginal DIC.

We begin, in Chapter 2, with a discussion of data involving environmental pollutants and the human menstrual cycle. Our analysis of these data, published in *Environment International* earlier this year, demonstrates that urinary biomarkers of environmental pollutants can predict important clinical endpoints related to the menstrual cycle. Moreover, we use a

marginalized definition of DIC to show that for each clinical endpoint considered, models involving these biomarkers are preferable to models involving only demographic information on the study participants. Our analysis substantiates the need for further investigation of environmental pollutant biomarkers in studies of human fertility, an area that has not been widely studied prior to our work.

Because data on environmental pollutant biomarkers are expensive to obtain and our collaborators were limited in their funding, the analysis we present in Chapter 2 involves methodological work to augment our observations that have biomarker data with other observations on the same individuals where this data is missing. This method, which we call shared parameter (SP) modeling and develop in the latter half of Chapter 2, uses a full model for the observations with biomarker data and a reduced model for the observations without this data. The method links these models by constraining them to share the same structure and coefficient values. The biomarker data are included in the full model, and the reduced model is augmented with an additional error term to mitigate the lack of biomarker data. Using our environmental epidemiology data, the SP modeling structure allows us to more precisely estimate participant-level random effects and coefficients for demographic covariates. Through simulation, we show that this method leads to better predictive accuracy than we would be able to achieve by using only the complete observations.

Chapter 3 presents a detailed introduction to the use of information criteria and related methods for statistical model selection. We focus on the DIC, and demonstrate how it is not well-defined for hierarchical models. Using a random effects model, we demonstrate that the complexity penalty for the DIC—a quantity called p_D—depends on how one defines the DIC relative to the random effects terms. We make the philosophical argument that the proper definition of DIC for hierarchical models is the marginalized DIC, where nuisance parameters from the hierarchical modeling are integrated out before DIC is computed. We

then show two ways this can be accomplished in the linear mixed model when random effects are normally distributed.

Obtaining a numerical approximation to the marginalized DIC is more difficult when generalized linear mixed models (GLMMs) are considered. Chapter 4 presents a new method for approximating the marginal density of a special case of GLMMs. Members of this special case of GLMMs are characterized by having "repeated exchangeable observations", and we therefore call them REO GLMMs. A GLMM has repeated exchangeable observations when observations within a cluster can be freely permuted without changing the resulting inferences. REO GLMMs are GLMMs for which covariates are constant for all observations within a cluster.

In the REO GLMM setting, we develop an approximation to the marginal density using Taylor's theorem and the complete-the-square formula. We prove that our approximation to the joint density for the data and the random effects converges almost surely to the true joint density for the data and the random effects. Through small-sample simulation, we then show that our approximation method gives joint density values close to the truth when more than 10 observations are available per cluster; and that our approximation to the marginal density gives p_D and DIC values close to the numerical approximations one would get using Gaussian quadrature to approximate the true marginal density. Finally, we simulate a model selection procedure to compare the behavior of a backwards stepwise algorithm using three different definitions of DIC as selection criterion: our approximation to the marginal DIC, our calculation of the joint DIC where random effects are treated as parameters, and the DIC statistic reported by OpenBUGS. We find that using our approximation to the marginal DIC reliably results in the selection of models more nearly matching the "true" model from which the simulation data were generated.

Chapter 5 extends our development of the approximation to non-REO GLMMs by introducing a Newton-Rhapson step. Using this, we derive a general form for the approximate

marginalization that allows us to consider time-varying covariates in longitudinal models as well as random slope models. We simulate new data including time-varying covariates and a cluster-specific random effect. We show that our approximation still gives joint density values close to the truth, and that it results in numerical approximations to p_D and DIC that match those obtained using the Gaussian quadrature. Using a simulated model selection procedure like the one in Chapter 4, we also show that selection based on our approximation to the marginal DIC results in models closer to the true generating model than selection based on the joint DIC or the DIC reported by OpenBUGS.

Chapter 2

Environmental Effects on Human Fertility and Shared Parameter Modeling

In this chapter, we discuss our work on an environmental epidemiology dataset. We begin by reviewing our previously published findings about the impact of environmental pollution on the human menstrual cycle. This work, conducted under the guidance of Drs. Ulrike Luderer and Wesley O. Johnson, has appeared in the peer-reviewed journal *Environment International* (Luderer et al. 2017). Portions are reprinted here with the permission of the publisher. Following this, we discuss in more detail a method we developed to analyze these data, which we refer to as shared parameter modeling. We created this method to deal with a complication involving the data in our study. Because of cost and funding issues, data on key covariates of interest were only available for half of our response observations. This forced us to ask how we could best make use of our available data, knowing that much of it was unable to directly address the central scientific question of our study.

We begin with work we have done to model response data when some covariate information is missing on many of the available response observations. We believe that a scientist should never throw away good data. In this chapter, we propose a method we call shared parameter (SP) modeling, which combines models using full covariate information and partial covariate information in a novel way. We begin by giving a practical example of this issue as it arises in data from our own research, and explain more generally the reasons scientists may encounter this problem. A technical definition follows, along with associated discussion of the assumptions of this method and how it relates to other methods. Finally, we use simulations to compare our method to alternative approaches scientists might use in this situation, and discuss the method's application in an analysis published earlier this year.

2.1 Environmental Effects on Human Fertility

Polycyclic aromatic hydrocarbons (PAHs) are ubiquitous environmental pollutants formed during incomplete combustion of organic materials such as wood, tobacco, fossil fuels, and food (Li et al. 2008; ATSDR 1995). Data from the National Health and Nutrition Examination Survey on concentrations of hydroxylated PAHs in the urine of representative samples of Americans show that essentially all Americans are exposed to PAHs (NHANES 2009). For non-smokers who do not consume grilled or roasted foods, air pollution is the largest source of exposure. Residents of urban areas have higher inhalation exposure to PAHs than do residents of rural areas (Menzie et al. 1992).

Many PAHs are mutagenic and carcinogenic (IARC 2010; ATSDR 1995; IARC 1983) and are potent ovarian toxicants and ovarian tumorigens in rodents. Neal, Zhu, and Foster (2008) measured PAHs in human serum and ovarian follicular fluid and showed that follicular fluid of women smokers had significantly elevated levels of the PAH benzo[a]pyrene (BaP) compared to follicular fluid of nonsmoking women. BaP inhibits growth, survival, and estradiol and

anti- Mllerian hormone secretion of cultured mouse secondary follicles (Neal, Zhu, Holloway, et al. 2007; Sadeu and Foster 2011).

PAHs are readily absorbed through the lungs, gut, and skin (ATSDR 1995). To exert toxicity, PAHs generally require metabolic activation through a process known as hydroxylation—a detoxification process by which organisms break down organic compounds into compounds that are more easily excreted. Hydroxylated PAHs (OH-PAHs) are PAHs that have had a hydroxyl group (an oxygen atom bonded to a hydrogen atom, thus the notation OH for hydroxyls) added to them. Biomonitoring of these hydroxylated PAH metabolites in urine provides an integrated measure of PAH exposure via multiple exposure routes (Li et al. 2008).

Although the altered reproductive function observed in women who smoke or who are exposed to environmental tobacco smoke (Harlow and Signorello 2000; Mattison et al. 1989) has been postulated to be due to exposure to PAHs in tobacco smoke, no studies have directly examined the associations between PAH exposure biomarkers and measures of hypothalamic-pituitary-ovarian axis function in women. We hypothesized that PAH exposure causes ovarian dysfunction manifested as altered urinary luteinizing hormone (LH) and estrogen metabolite profiles and even anovulatory menstrual cycles.

2.1.1 Methods

2.1.1.1 Study participants

Study participants were recruited for a pilot study to lay the groundwork for a subsequent larger, adjunct study to the planned National Children's Study (NCS). The purpose of the pilot study was to test the feasibility of recruiting women who were not intending to become pregnant and not using hormonal contraception for a study of the association between urinary

PAH metabolites measured once per menstrual cycle and urinary reproductive hormone concentrations measured daily using a microelectronic dipstick monitor for six menstrual cycles.

Eligible women were between the ages of 1844, residing in Orange County California, who were not pregnant, currently not planning to conceive, not using hormonal contraception, did not have a history of surgical sterilization, treatment with antineoplastic drugs or radiation therapy to the pelvis and did not have conditions known to cause infertility by mechanisms other than ovarian failure (pelvic inflammatory disease, endometriosis). Women who had recently been pregnant or breastfeeding were asked to delay starting the study until they had one full menstrual cycle after the birth if not breastfeeding or after they stopped breastfeeding.

Initially eligible women were identified and recruited by door-to-door contact in the home, with follow-up telephone and email contact by study staff. Subsequently, when the NCS shifted to other recruitment strategies, eligible women for the present study were recruited at public events such as health fairs at universities and colleges, work places and events sponsored by community groups. The current study population is thus a convenience sample. Baseline study visits and sample collection occurred between October 2010 and July 2012.

2.1.1.2 Baseline study visit

After completing informed consent, participants were instructed to go to one of the two Orange County locations of the UC Irvine Institute for Clinical and Translational Science (ICTS) for their baseline visit 59 days prior to their next menses onset. Study staff administered the standardized NCS preconception questionnaire to obtain information about the participant's demographics, medical history, reproductive history, tobacco smoke exposure, exercise history, occupational history, residential history, housing characteristics, use

of chemicals (e.g. cleaning agents, pesticides) in and around the home and yard, and pets. Questions relating to tobacco, alcohol, and illicit drug use were taken from the NCS First Trimester Maternal In-Person Questionnaire. We refer to these variables hereafter as baseline covariates.

ICTS nursing staff measured height, weight, and blood pressure and collected a blood sample. Study staff gave participants urinary hormone monitors, urine collection kits, and home pregnancy tests and provided instruction in how to use them.

2.1.1.3 Urinary reproductive endocrine testing

Participants were given a Clearblue Easy Fertility Monitor (Swiss Precision Diagnostics, Bedford, UK) and instructed to perform daily urinary dipstick tests to measure estrone 3-glucuronide ($E_1$3G) and LH beginning on the first day of their next menstrual cycle. The Clearblue Easy Fertility Monitor measures daily urinary $E_1$3G and LH without any need for collection or storage of urine samples. For each menstrual cycle, the participant pushes the monitor M button if she had onset of menstrual bleeding during the preceding 24h (cycle day one). The monitors request daily test sticks starting on cycle day 6 and continuing until an LH surge has been identified or until 20 days have passed, whichever is first. In subsequent cycles, the monitor requests tests based on the timing of the LH surges in prior cycles. To conduct a test, the participant holds a disposable test stick in the urine stream of the first morning void for 3s. The monitor displays a fertility status (low, high, or peak), derived by proprietary algorithmic interpretations of the LH and $E_1$3G concentrations. Participants were also instructed to perform a human chorionic gonadotropin Clearblue Easy Pregnancy Test (Swiss Precision Diagnostics) if their menses onset did not begin within 10 days of their expected date. If the pregnancy test was positive, the participants were removed from the current study.

The list below describes endpoints that were analyzed in this study. These endpoints are all summaries of the data obtained from the Clearblue Easy Fertility Monitors. LH surge onset and day of $E_1$3G peak (not listed) were used to calculate other endpoints and were not further analyzed. Menses onset for each cycle was determined from the monitor output and/or the daily diary. Analysis indicated that almost all quantitative endocrine endpoints had approximately normal distributions, with the sole exception of menstrual cycle length, which was heavily right-skewed.

Cycle Length	Number of days from first day of menstrual bleeding through the day before next onset of menstrual bleeding.
Follicular Phase Length	Number of days from first day of menstrual bleeding to the day of the LH surge onset.
Ovulatory Status	Ovulatory cycle = Has a defined LH surge onset; Anovulatory cycle = No LH surge onset for cycles with no missing LH values from cycle day 11 through the 9th day before the next menses onset. Must have start and end menses; Indeterminate cycle = Cycles that are neither ovulatory nor anovulatory
Follicular LH	Mean for all days before LH surge onset, or cycle days 6 through 10 for cycles without an LH surge. Recorded as missing if < 3 values.
Highest LH	Highest LH value of the cycle that is $\geq$2.5-fold above the LH surge onset baseline and $\leq$4 days after the LH surge onset.
Peak LH	Same as highest LH surge level except it is recorded as missing if there is a missing value on an adjacent day.

Follicular $E_1$3G	Mean for cycle day 1 through day -2 from LH surge onset, or for $E_1$3G values on days 6 through 10 for cycles with no LH surge onset. Recorded as missing if < 3 values.
Periovulatory $E_1$3G	Mean for 7 days centered around the day of the LH surge onset. Recorded as missing if < 3 values.
$E_1$3G Slope	Slope for 3 days prior to day of $E_1$3G peak. Recorded as missing if the first or third value is missing.

2.1.1.4 Measurement of urinary hydroxylated PAH metabolites

Participants were instructed to collect one urine sample on the 10^{th} day after menses onset into a polypropylene beaker and to pour 10 mL of that sample from the beaker into each of 4 cryogenic polypropylene vials to be stored in their home freezer until picked up by study staff every month. Hydroxylated PAH (OH-PAH) metabolites were measured in the urine samples as biomarkers of PAH exposure. We chose cycle day 10 because it is approximately in the middle of the follicular phase days when the monitor requests test sticks.

The OH-PAH compounds measured in this study included fluorene ($_2FLUO$, $_3FLUO$, $_9FLUO$), phenanthrene ($_1PHEN$, $_2PHEN$, $_3PHEN$), naphthalene ($_1NAP$, $_2NAP$), and pyrene ($_1PYR$). The numbering associated with these OH-PAHs refers to the bind point in the PAH where the hydroxyl group connects. Linguistically $_1NAP$ should be read as "1-hydroxy-naphthalene", with the other metabolites read similarly.

Urinary OH-PAH metabolites were measured for each subject for two cycles (3 participants) or three cycles (48 participants) during the testing period by the California Department of Public Health Environmental Health Laboratory using procedures developed by the Centers for Disease Control and Prevention NHANES study (Li et al. 2008; Romanoff et al. 2006). If

a participant only had collected two or three cycles of urinary hormone data, then OH-PAHs were measured in all cycles. If they had collected urinary hormone data for more than three cycles, then the first three cycles with the fewest missing days of urinary hormone data were selected for OH-PAH measurements.

Complications caused by only having OH-PAH measurements for 2-3 cycles per woman (approximately half of our observations) will be discussed further in Section 2.2. Note that the missingness paradigm for these data is "missing completely at random" (MCAR). Because 2-3 observations are available on each woman, missingness is not related to the individual women—or, thus, to their baseline covariates which are constant throughout the study. We assume further that missingness is not related to PAH measures or response measures. We can think of little reason why a woman's daily hormone levels or monthly exposure to environmental pollutants would impact which cycles had the fewest missing days of urinary hormone data.

A small percentage of analytes were below the limits of detection (LOD) of the assay. These were set to $LOD/\sqrt{2}$ (Ogden 2010).

2.1.1.5 Statistical analysis

Descriptive statistics (arithmetic means, standard deviations and geometric means for continuous variables; and percentages in each group for categorical variables) were calculated for demographic variables (Tables 2.1 and 2.2), endocrine endpoints (Table 2.3) and OH-PAH concentrations (Table 2.4). To assess the collinearity among OH-PAH metabolite concentrations within the same urine sample, we calculated pairwise Pearson correlations (Table 2.5).

Observed metabolite concentrations were right-skewed, but were approximately normal after a log-transformation was applied. These log transformed values were then standardized by

subtracting the corresponding sample mean and dividing by the corresponding standard deviation so that each new transformed variable has sample mean zero and sample standard deviation one.

Then, in part because concentrations of metabolites of phenanthrene ($_1PHEN$, $_2PHEN$, $_3PHEN$) and of fluorene ($_2FLUO$, $_3FLUO$, $_9FLUO$) can be collinear—see Table 2.5—additional transformations were considered for each. One of these transformations, $PHENs_{ij}$, is a standardized average of the three standardized phenanthrene variables for each woman i and time j combination. Designating $_1PHEN$, $_2PHEN$, and $_3PHEN$ as the standardized log-transformed concentrations of 1-, 2-, and 3-hydroxy phenanthrene respectively, then this variable is defined as

$$PHENs_{ij} = c_0(_1PHEN_{ij} + _2PHEN_{ij} + _3PHEN_{ij}),$$

where c_0 is a variance standardization constant. The variable $FLUOs$ is defined analogously. We define three additional transformations of the phenanthrene metabolites: ^d_1PHEN, ^d_2PHEN, and ^d_3PHEN. These are standardized differences between the concentrations of the individual phenanthrene metabolites and the overall average $PHENs$. The formula for ^d_1PHEN is the contrast

$$\begin{aligned} {}^d_1PHEN_{ij} &= c_1(2\,{}_1PHEN_{ij} - {}_2PHEN_{ij} - {}_3PHEN_{ij}) \\ &= 3c_1({}_1PHEN_{ij} - \frac{1}{3c_0}PHENs_{ij}) \\ &= 3c_1({}_1PHEN_{ij} - \overline{PHEN}_{ij}), \end{aligned}$$

where $\overline{PHEN}_{ij} = (_1PHEN_{ij} + _2PHEN_{ij} + _3PHEN_{ij})/3$ and c_1 is a standardizing constant. This is simply a measure of the difference between the concentration of 1-phenanthrene and the average concentration across all phenanthrene metabolites, in terms of standard

concentration scores. ${}^{d}_{2}PHEN$ and ${}^{d}_{3}PHEN$ are defined analogously as contrasts for ${}_{2}PHEN$ and ${}_{3}PHEN$ respectively. Similar fluorene contrasts are also defined analogously.

These transformations are intended to mitigate collinearity among these measures. They also allow for an "overall effect" measure of phenanthrene (for woman i at time j), and the three contrasts that are based on the degree to which a woman is above/below her average phenanthrene isomers at time j. That is, they allow us to study an "isomer specific" effect that reflects the degree to which one particular isomer does or does not match the overall level characterized by the combined covariate (for each woman-time combination).

The two naphthalene metabolites were not handled similarly because: (1) Table 2.5 shows that there is very little collinearity between concentrations of these metabolites, and (2) it is known that ${}_{1}NAP$ and ${}_{2}NAP$ result from the metabolizations of substantially different environmental compounds (Hill et al. 1995).

To investigate the role of OH-PAH concentrations in predicting the aforementioned endocrine endpoints, we performed a two-stage procedure. Models for all endpoints were constructed in the same way using a Bayesian mixed modeling approach with subject-specific random effects allowing each participant to act as a baseline for her set of measurements. In the first stage we built a model involving only non-PAH baseline covariates, using a backwards stepwise algorithm to select a parsimonious model that fit the data, explained below.

The initial baseline model included the covariates age, race, educational attainment, stress, body-mass index (BMI), alcohol use, caffeine intake, and measures of how many minutes the participant walked and engaged in vigorous physical activity each week. When covariate data were missing, which happened on 0-3 individuals per baseline covariate, values were imputed through a Bayesian modeling approach (Daniels and Hogan 2008, Ch.6). Existing covariate/endpoint combinations were used impute the missing values in the MCMC algorithm. As a result of modeling the missing values, uncertainty about their actual values is

incorporated into the analysis. Our modeling approach for follicular phase length involves standard linear modeling assumptions. Thus, conditional on the covariates and the random effects, we assume independence among the response values and normal errors with constant variance. Random effects are modeled with a normal distribution with mean zero and a random effects variance.

Figure 2.1 shows posterior density estimates for all baseline regression coefficients in the first phase of modeling for follicular phase length. Observe that several density plots are centered close to zero. The corresponding covariates are candidates for removal from the model. At the same time, other plots are concentrated below zero, indicating high posterior probability of a negative association, while yet other plots indicate a high posterior probability of a positive association, for those variables, respectively. Our backwards stepwise approach involved considering the posterior probabilities of coefficient values being above zero for each covariate (or below zero if the covariate estimate is negative). We removed the covariate with corresponding proportion closest to 0.5 (i.e. an even split between positive and negative coefficient estimates across all simulated iteratesa good proxy for measuring how well coefficients cluster away from zero). Next, a new regression model was fit with the remaining covariates and the removal procedure repeated until all remaining covariates showed non-zero coefficient proportions greater than 0.85. In the case of the Figure 2.1 example, the walking coefficients ($\beta_{W_1}, \beta_{W_2}, \beta_{W_3}$) show the smallest proportion of values away from zero[1], and walking was removed from the model as a covariate. Based on standard epidemiological practice, age and race/ethnicity were forced into the model as known factors of interest and were not subject to the removal procedure discussed.

We designated the model resulting from this stepwise procedure as our baseline model for each particular endpoint. Our primary research interest is to assess the effect of PAHs on

[1]When dealing with covariates with multiple levels, such as walking, we based our decisions on the coefficient whose posterior probability of being above zero was furthest from 0.5. If this coefficient had a probability of being above zero nearer 0.5 than the coefficients for every other variable, the covariate was dropped.

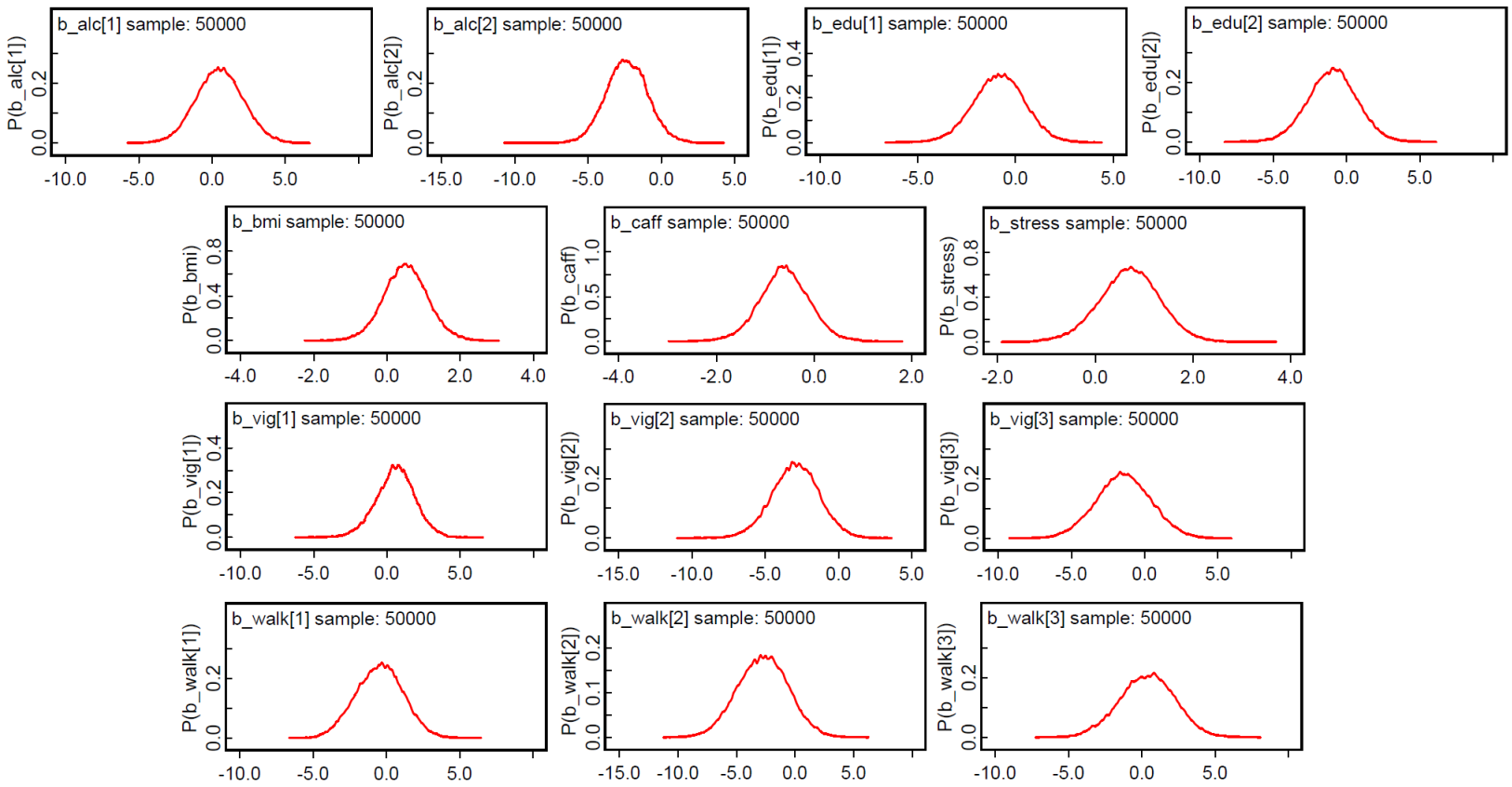

Figure 2.1: Posterior density estimates for regression coefficients on follicular phase length. These are taken from the starting model for backwards stepwise selection in the first stage of our model-finding procedure (i.e. the stage in which we find the baseline model). The covariate to be removed is the one whose coefficient(s) have posterior densities nearest to zero.

endocrine endpoints, and so our goal in these analyses was to determine whether OH-PAHs can provide any incremental benefit in modeling those endpoints beyond the fit we obtained with the baseline model. Observing such an incremental benefit provides evidence not only that PAHs relate to endocrine endpoints, but also that PAHs contribute unique information about those endpoints that we cannot capture with our baseline covariates.

The second stage model then began from the selected baseline covariate model, adding all of the OH-PAH variables and an indicator variable denoting whether the participant was a smoker to that model. Smoking was only added at this stage because we believe PAHs represent the mechanism of action for smoking on these covariates. Because OH-PAH information was only available for 23 cycles per participant, we used a shared parameter (SP) modeling strategy to take full advantage of our available endpoint data. This approach is described in more detail in Section 2.2. As in the first stage, we used a backwards selection procedure to pick a parsimonious model that fit the data well with any additional terms that had been selected. Our primary research interest was to assess the effect of PAHs on endocrine endpoints, and so our goal in these analyses was to determine whether OH-PAHs could provide any incremental benefit in modeling those endpoints beyond the fit we obtained with the baseline model.

Finally, we compared the deviance information criterion (DIC, a common Bayesian measure of model fit) for the baseline and final models. A known concern with DIC is that the value of the criterion depends on the collection of parameters being evaluated (Celeux et al. 2006; D. J. Spiegelhalter et al. 2002). To address this issue, numerical approximations to DIC for these models were based on the marginal model involving only parameters of interest. This is an application of our "postprocessing marginalization" method for obtaining DIC in linear mixed models, discussed in Section 3.4.1.2. Generally, a decrease of 3 or more in DIC is suggestive that a lower-scoring model is preferred to a higher-scoring model (D. J. Spiegelhalter et al. 2002). In this way, we addressed a primary goal of this paper

by assessing the added benefit of OH-PAH terms over and above the model involving only baseline characteristics.

When needed, estimates of regression coefficients were their posterior means. These were numerically approximated using Markov Chain Monte Carlo (MCMC) simulation (R. R. Christensen, Johnson, et al. 2010, Ch.6). This allowed us to directly examine the posterior probability that each covariate in the model was associated with a nonzero regression coefficient by determining the proportion of simulated coefficient values that were above zero. A covariate with coefficient values clustering around zero is deemed to be unimportant in terms of its contribution to modeling follicular phase length, whereas a covariate whose coefficient values are almost all positive (or almost all negative), corresponding to high posterior probability that the coefficient is indeed positive (or negative), demonstrates a statistically important contribution.

Analyses were performed in WinBUGS using a Gibbs sampling procedure with 50,000 iterations following a 5000-iteration burn-in period to ensure convergence of the MC. The choice of 50,000 iterations reduced the simulation error in numerical approximations of the aforementioned proportions.

Convergence of the MC was assessed by running separate chains on multiple endpoints during early stages of the model selection procedure. For each model so examined, all chains showed rapid convergence to the same posterior distribution. Early models were also run with a set of three possible priors for the regression coefficients: a skeptical N(0,9) prior that presupposed that regression coefficients would be near 0, a diffuse N(0,10000) prior that allowed regression coefficients to take a wide range of values, and a middle N(0,100) prior. We found that results of our stepwise procedure were not particularly sensitive to our choice of prior, though the skeptical prior was informative enough to depress coefficient estimates.

Final analyses were all conducted using the N(0,100) prior on regression coefficients, which we considered to be the most reasonable choice in terms of the magnitude of regression coefficients we would expect to see for standardized covariates and for endpoints on the scales reported (see Table 2.3 for information on the scale of endpoint values). Error variances were modeled with a Gamma(0.001,0.001) prior on the precision (the inverse of the variance). Random effects variances were modeled by putting a Uniform(0,50) prior on the standard deviation.

2.1.2 Results

2.1.2.1 Demographics and participant characteristics

Three participants were eliminated from the analyses because they did not collect any urine for PAH measurements (one woman) or they had performed so few urine hormone test sticks that the no endocrine endpoints could be calculated (two women). Demographic and other characteristics of the 51 remaining participants are shown in Tables 2.1 and 2.2. The mean age of the participants was 29.9 years. Non-Hispanic white women made up the largest racial/ethnic group among the participants, followed by Asian and Hispanic, in that order. More than 92% of the participants had some education beyond high school graduation, with 21.6% having a graduate degree. The mean BMI was 24.9.More than a third of the participants had engaged in more than 2 h of vigorous exercise,while nearly 24% had not engaged in any vigorous exercise, during the week prior to the baseline interview. Forty-one percent of the participants reported walking ¿4h during the week prior to the baseline interview. Fewer than 12% of the participants smoked, and 23.5% reported drinking alcoholic beverages on 2 or more days per week.

Characteristic	Mean ± SD	Range	N
Age at baseline	29.9 ± 6.5	$[18, 44]$	51
Height (m)	1.63 ± 0.07	$[1.50, 1.83]$	48
Weight (kg)	66.6 ± 15.7	$[49.2, 124.9]$	49
BMI	24.9 ± 5.3	$[18.6, 42.6]$	48

Table 2.1: Continuous characteristics of study participants.

2.1.2.2 Menstrual cycle endpoints

We had monitor data for 305 menstrual cycles from the 51 participants. Of these, 150 cycles had OH-PAH measurements. Representative menstrual cycle LH and $E_1$3G concentrations for a participant with regular cycles with clearly defined LH peaks are shown in Figure 2.2a, while representative data for a participant with some anovulatory cycles are shown in Fig. 2.2b. Both of these participants missed very few days of sampling. In contrast, some participants had many missing days. Of 4726 potential test days recorded by the monitors, tests were not performed on 1168 days. Most of these were because the participant did not turn the monitor on at all that day or turned it on outside of the test window; 10 participants had one to several cycles when they took a hiatus for travel or other reasons and resumed thereafter.

Endocrine endpoints summarized from these data are presented in Table 2.3. This table also shows for how many cycles each endpoint could be calculated. Cycle length was calculable for the largest number of cycles, 297 of 305 cycles. The mean follicular phase LH and $E_1$3G concentrations were calculable for the next largest numbers of cycles because these variables do not depend on having identified a mid-cycle LH surge.

Characteristic	N	Percent
Ethnicity/Race		
Hispanic	11	21.6%
Non-Hispanic White	22	43.1%
Non-Hispanic Asian	14	27.5%
Non-Hispanic Black	4	7.8%
Education – highest level attained		
High school diploma or less	3	5.9%
Some college or vocational	20	39.2%
Bachelor's degree	16	31.4%
Graduate degree	11	21.6%
Family income/year		
$< \$20,000$	4	7.8%
$\geq \$20,000 < \$50,000$	12	23.5%
$\geq \$50,000 < \$75,000$	10	19.6%
$\geq \$75,000 < \$100,000$	9	17.6%
$\geq \$100,000$	9	17.6%
Current smoking		
Yes	6	11.8%
No	45	88.2%
Alcohol		
< 1 day/month or never	21	41.2%
$1 - 4$ days/month	16	31.4%
≥ 2 days/wk	12	23.5%
Caffeinated beverages (drink regularly)		
Coffee	28	54.9%
Tea	21	41.2%
Soda	10	19.6%
Energy drinks	2	3.9%
Vigorous exercise (m in last 7 days)		
0	12	23.5%
$> 0 \leq 120$	22	43.1%
$> 120 \leq 390$	10	19.6%
> 390	7	13.7%
Walking (h in last 7 days)		
≤ 1	9	17.6%
$> 1 \leq 4$	21	41.2%
$> 4 \leq 7$	8	15.7%
> 7	13	25.5%

Table 2.2: Discrete characteristics of study participants.

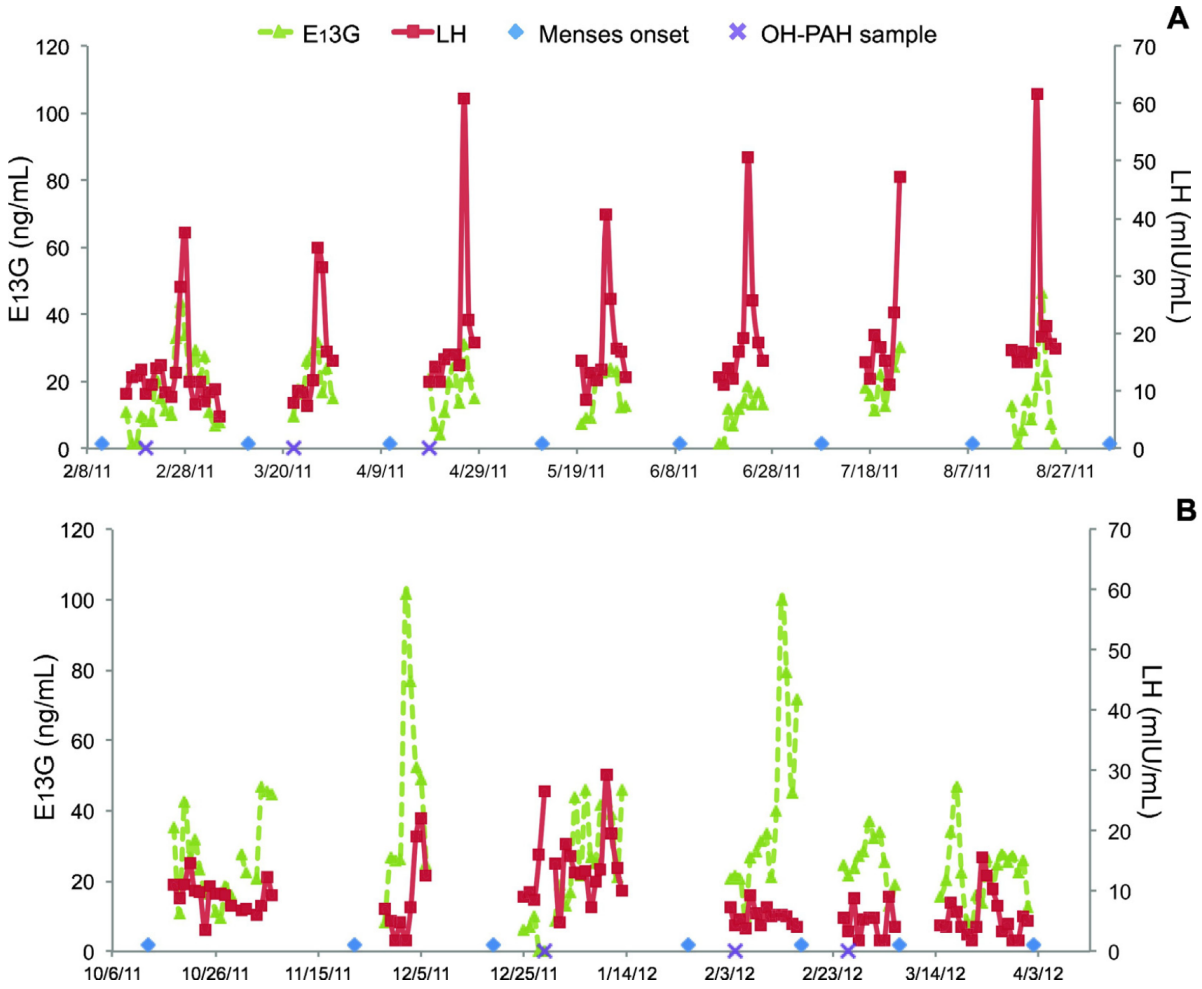

Figure 2.2: Representative urinary LH and $E_1$3G concentrations for two participants across multiple menstrual cycles. A) A participant with regular cycles with clearly defined LH peaks. B) A participant with several ostensibly anovulatory cycles without clear LH peaks. Diamonds indicate the onset of menses. Days for which urinary OH-PAHs were measured are indicated by an X.

Endpoint variable	Mean (SD)	Geometric mean	N	n
Cycle length (days)	33.2 (10.7)	32.0	51	297
Follicular phase length (days)	16.3 (3.3)	16.0	47	199
Follicular LH (mIU/mL)	9.1 (2.0)	8.9	59	233
Highest LH (mIU/mL)	36.6 (10.7)	35.3	47	199
Peak LH (mIU/mL)	37.9 (12.1)	36.1	45	132
Follicular $E_1$3G (ng/ml)	20.2 (7.5)	18.7	50	219
Periovulatory $E_1$3G (ng/ml)	31.1 (10.3)	29.4	47	191
$E_1$3G slope	3.2 (3.7)	3.2	49	178

Table 2.3: Means, standard deviations, and geometric means of continuous menstrual cycle endpoint variables averaged over within-patient averages. N refers to the number of women for whom there was at least one cycle of data for that variable. n refers to the total number of cycles for which each endpoint could be calculated, out of the 305 cycles observed.

2.1.2.3 PAH exposures

OH-PAHs were measured around menstrual cycle day 10 for three cycles (48 participants) or two cycles (3 participants) per participant. Table 2.4 shows the arithmetic and geometric mean concentrations of urinary OH-PAH metabolites. All participants had detectable concentrations of each of the measured OH-PAHs in at least one urine sample, which is consistent with ubiquitous exposure to PAHs.

OH-PAH	Mean	SD	Geometric mean
1-hydroxy-naphthalene	1785	1245	1387
2-hydroxy-naphthalene	6192	5037	4965
1-hydroxy-pyrene	*107*	*64*	*95*
1-hydroxy-phenanthrene	119	60	107
2-hydroxy-phenanthrene	33	14	31
3-hydroxy-phenanthrene	59	23	55
2-hydroxy-fluorene	*209*	*121*	*184*
3-hydroxy-fluorene	*106*	*64*	*91*
9-hydroxy-fluorene	*296*	*259*	*240*

Table 2.4: Means, standard deviations, and geometric means of OH-PAH concentrations averaged over within-patient averages. Reported concentrations are all creatinine adjusted (ng/(g creat).

Pairwise correlations among the different OH-PAHs within urine samples are shown in Table 2.5. All of the measured metabolites tended to have moderate correlation with each other, with higher correlations for metabolites of the same parent compound. $_2NAP$ was the exception; it had relatively low correlation with the other metabolites including $_1NAP$. The moderately high collinearity among metabolites of fluorene and among metabolites of phenanthrene demonstrated in Table 2.5 provides some justification for the combination and difference-score forms we used for these metabolites in the endocrine endpoint analyses.

2.1.2.4 Associations between OH-PAH metabolites and endocrine endpoints

Tables 2.6 and 2.7 present (1) the final OH-PAH models for each endocrine endpoint, (2) PAH coefficient estimates, (3) the proportion of positive coefficients from the MCMC, (4) a list of baseline covariates identified by the stepwise modeling procedure for which the models have been adjusted, and (5) DIC scores for comparing the baseline and final models. Creatinine-adjusted OH-PAH measurements were used for all models displayed. The models for all endpoints were refit with unadjusted OH-PAH measurements, and the results did not differ in any appreciable respect, and thus are not presented here.

For all eight endpoints considered, addition of OH-PAHs or smoking status improved the model fit (decreased DIC by > 3) beyond that provided by the covariates in the baseline models. Smoking status played a role in modeling the three $E_1$3G endpoints, but directly measured OH-PAHs were selected into all eight models even when smoking status was accounted for.

Because understanding what these models are saying can be difficult, we provide graphical interpretations using the follicular phase length (Figure 2.3) and highest LH (Figure 2.4) endpoints. We focus first on Figure 2.3. Consider a hypothetical participant with urinary $_1PYR$ concentration of 85 ng/g Cr for her first menstrual cycle, and 280 ng/g Cr for her second

OH-PAH	$_1NAP$	$_2NAP$	$_1PYR$	$_1PHEN$	$_2PHEN$	$_3PHEN$	$_2FLUO$	$_3FLUO$	$_9FLUO$
$_1$NAP	1.000	0.105	0.234	0.429	0.285	0.352	0.374	0.446	0.424
$_2$NAP		1.000	0.163	-0.038	0.030	-0.011	0.072	0.067	-0.113
$_1PYR$			*1.000*	*0.597*	*0.574*	*0.654*	*0.385*	*0.465*	*0.379*
$_1$PHEN				1.000	0.738	0.738	0.635	0.565	0.655
$_2$PHEN					1.000	0.764	0.573	0.536	0.615
$_3$PHEN						1.000	0.638	0.683	0.631
$_2FLUO$							*1.000*	*0.887*	*0.470*
$_3FLUO$								*1.000*	*0.459*
$_9FLUO$									*1.000*

Table 2.5: Pearson correlations among concurrent log-transformed concentrations for all 151 urine samples for which OH-PAHs were measured.

Endpoint	DIC (w/o PAHs)	DIC (w/ PAHs)	
Cycle length	1211.5	1203.3	
Baseline	Race + Age + Walking + Vigorous Activity		
PAHs	${}^{d}_{9}FLUO$	$\beta = -0.507$	$\Pr[\beta < 0] = 0.960$
	${}^{d}_{2}PHEN$	$\beta = 0.323$	$\Pr[\beta > 0] = 0.889$
	${}_{1}PYR$	$\beta = 0.728$	$\Pr[\beta > 0] = 0.994$
Folli. phase length	977.9	961.6	
Baseline	Race + Age + Stress + Vigorous Activity		
PAHs	${}^{d}_{3}FLUO$	$\beta = 1.040$	$\Pr[\beta > 0] = 0.999$
	${}_{1}PYR$	$\beta = 0.578$	$\Pr[\beta > 0] = 0.982$
Follicular $E_1$3G	1428.7	1420.0	
Baseline	Race + Age + Caffeine[a] + Walking		
PAHs	${}^{d}_{2}FLUO$	$\beta = 1.325$	$\Pr[\beta > 0] = 0.979$
	${}^{d}_{3}FLUO$	$\beta = -1.854$	$\Pr[\beta < 0] = 0.999$
	${}_{1}NAP$	$\beta = 0.830$	$\Pr[\beta > 0] = 0.924$
	${}_{2}NAP$	$\beta = -0.680$	$\Pr[\beta < 0] = 0.878$
	${}^{d}_{2}PHEN$	$\beta = -0.613$	$\Pr[\beta < 0] = 0.870$
	Smoking	$\beta = 4.975$	$\Pr[\beta > 0] = 0.940$
Periovulatory $E_1$3G	1322.9	1318.8	
Baseline	Race + Age + Caffeine + Alcohol + Vigorous Activity + Education		
PAHs	$FLUOs$	$\beta = 1.595$	$\Pr[\beta > 0] = 0.920$
	${}_{2}NAP$	$\beta = -1.381$	$\Pr[\beta < 0] = 0.954$
	${}_{1}PYR$	$\beta = -0.985$	$\Pr[\beta < 0] = 0.853$
	Smoking	$\beta = 8.585$	$\Pr[\beta > 0] = 0.969$
$E_1$3G Slope	1097.7	1094.4	
Baseline	Race + Age + Stress[a] + Alcohol[a] + Walking + Vigorous Activity		
PAHs	${}_{1}NAP$	$\beta = 0.967$	$\Pr[\beta > 0] = 0.954$
	${}_{2}NAP$	$\beta = 0.815$	$\Pr[\beta > 0] = 0.944$
	${}^{d}_{1}PHEN$	$\beta = -0.952$	$\Pr[\beta < 0] = 0.939$
	Smoking	$\beta = 1.809$	$\Pr[\beta > 0] = 0.878$

Table 2.6: Baseline and PAH models for cycle length, follicular phase length, and $E_1$3G endpoints. Baseline covariates marked with an [a] were included in the baseline model but dropped in the PAH model under our selection criteria.

Endpoint	DIC (w/o PAHs)	DIC (w/ PAHs)	
Follicular LH	1072.6	1068.3	
Baseline	Race + Age + Caffeine[a] + BMI + Alcohol + Vigorous Activity		
PAHs	${}_2NAP$	$\beta = 0.446$	$\Pr[\beta > 0] = 0.962$
	$PHENs$	$\beta = -0.963$	$\Pr[\beta < 0] = 0.996$
	${}_1PYR$	$\beta = 0.548$	$\Pr[\beta > 0] = 0.945$
Highest LH	1563.2	1557.3	
Baseline	Race + Age + Caffeine[a,b] + BMI[b] + Alcohol[a] + Walking[a] + Education[a]		
PAHs	${}^d_2FLUO$	$\beta = 2.057$	$\Pr[\beta > 0] = 0.932$
	${}^d_3FLUO$	$\beta = -2.000$	$\Pr[\beta < 0] = 0.941$
	$FLUOs$	$\beta = 5.308$	$\Pr[\beta > 0] = 0.983$
	${}_1NAP$	$\beta = -1.977$	$\Pr[\beta < 0] = 0.924$
	${}_2NAP$	$\beta = -3.143$	$\Pr[\beta < 0] = 0.936$
Peak LH	1055.7	1035.8	
Baseline	Race + Age + Caffeine[a,b] + BMI[a,b] + Alcohol[a] + Walking[a] + Education[a]		
PAHs	${}^d_2FLUO$	$\beta = 2.755$	$\Pr[\beta > 0] = 0.949$
	${}^d_3FLUO$	$\beta = -2.988$	$\Pr[\beta < 0] = 0.981$
	FLUOs	$\beta = 7.262$	$\Pr[\beta > 0] = 0.992$
	${}_1NAP$	$\beta = -3.735$	$\Pr[\beta < 0] = 0.981$
	${}_2NAP$	$\beta = -1.823$	$\Pr[\beta < 0] = 0.893$
	${}^d_2PHEN$	$\beta = -2.371$	$\Pr[\beta < 0] = 0.950$
	$PHENs$	$\beta = -4.280$	$\Pr[\beta < 0] = 0.967$

Table 2.7: Baseline and PAH models for LH endpoints. Baseline covariates marked with an [a] were included in the baseline model but dropped in the PAH model under our selection criteria. Baseline covariates marked with a [b] were forced into the model for Peak LH because they were also present in the model for Highest LH, which uses the same data.

cycle. A ${}_1PYR$ concentration of 85 ng/g Cr is approximately the median concentration in our dataset, and a concentration of 280 ng/g Cr reflects a value two standard deviations higher on the log scale. Then the expected increase in follicular phase length is 1.156 days longer in the second cycle than in the first (Fig. 2). In addition, the graphs in Figure 2.3 show that a shift in ${}_2FLUO$ or ${}_9FLUO$ from their median to their 75th percentile concentrations is associated with approximately a half day decrease in follicular phase length, while a similar shift in ${}_3FLUO$ is associated with about a one day increase.

For a sense of the size of the effects in Figure 2.3, based on our data we find that given identical covariate values, the difference between the interquartile range for a woman's follicular phase length spans 3.3 days. Similarly, the interquartile range for follicular phase length among all women spans about 3.7 days. A predicted difference of one to two days based on OH-PAH concentration levels is less than the expected amount of cycle-to-cycle variability within women, but it is still a noteworthy effect.

Highest LH and Peak LH were calculated from the same data when a surge was identified with no missing data. Peak LH was not calculated when the peak of the LH surge could not be confirmed due to missing data. Results for both endpoints are similar, as we would hope, and show roles for all naphthalene and fluorene metabolites studied, as well as some naphthalene and phenanthrene metabolites. Of particular note, the coefficients corresponding to ${}^d_2FLUO$ and ${}^d_3FLUO$ have opposite signs and similar magnitudes. Recall that these covariates represent the difference between their specific metabolite and the overall shared level of the fluorene compound across all metabolites. What we appear to see here, then, is that highest LH increases when the ${}^d_2FLUO$ metabolite accounts for a greater proportion of the total fluorene in a subject's urine than the ${}^d_3FLUO$ metabolite. We also see that increasing levels of fluorene overall are associated with an increase in the highest LH level. Figure 2.4 provides a graphical representation of the effects of changes in ${}_1NAP$, ${}^d_2FLUO$, and ${}^d_3FLUO$ on highest LH. We see from the graphs that highest LH decreases with increasing ${}_1NAP$. A

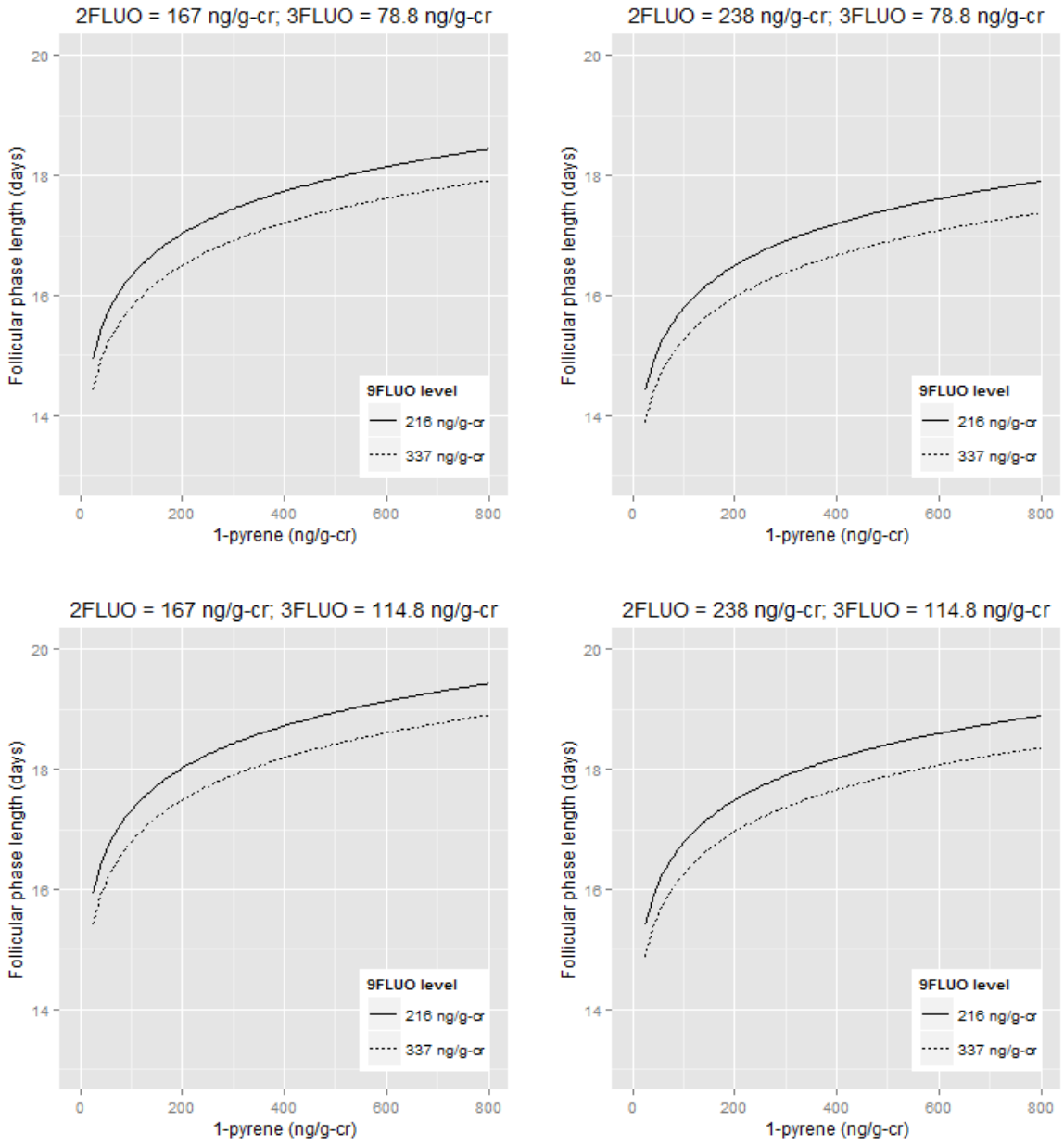

Figure 2.3: Graphical representation of changes in follicular phase length with changes in pyrene and fluorene metabolites derived from the final model. All four graphs show the changes in follicular phase length with increasing urinary concentration of ${}_1PYR$ at two different concentrations of ${}_9FLUO$, the study median concentration (solid line) and the 75th percentile concentration (dotted line). ${}_2FLUO$ was held constant at its study median concentration in the two graphs on the left and at its 75th percentile in the two graphs on the right. ${}_3FLUO$ was held constant at its study median concentration in the two upper graphs and at its 75th percentile in the two lower graphs. The graphs show that follicular phase length increases with increasing ${}_1PYR$. A shift in ${}_2FLUO$ or ${}_9FLUO$ from the median to the 75th percentile results in approximately a half day decrease in follicular phase length, while a similar shift in ${}_3FLUO$ is associated with a one day increase.

shift in ${}_2FLUO$ or ${}_9FLUO$ from their medians to their 75th percentile concentrations results in approximately a 5 and 1 mIU/mL increase in highest LH, respectively, while a similar shift in ${}_3FLUO$ is associated with about a 3 mIU/mL decrease in highest LH.

For the three follicular phase $E_1$3G endpoints, we observed positive associations with smoking. On average, the slope of the estradiol rise is steeper in smokers than the slope in non-smokers. In practical terms, for a white woman of average age (in the study) who doesn't walk or exercise much, the slope for nonsmokers shows an increase in $E_1$3G of about 2 ng/mL per day, and the slope for smokers shows a daily increase of about 4 ng/mL. $E_1$3G slope also increases with ${}_1NAP$ and ${}_2NAP$ and decreases when ${}_1PHEN$ is higher than the other standardized phenanthrene metabolite concentrations. On average, the periovulatory $E_1$3G concentrations are higher for women who smoke by nearly 9 ng/mL. A white woman of average age who doesn't exercise much, doesn't have a college degree, doesn't drink caffeinated beverages, and doesn't drink alcohol will on average have a periovulatory $E_1$3G of about 20 ng/mL if she doesn't smoke and 28.5 ng/mL if she does smoke. Periovulatory $E_1$3G also increases with fluorene metabolites and decreases with ${}_2NAP$ and ${}_1PYR$. On average, follicular phase estradiol is higher for smokers, with our model predicting a 5-unit increase in $E_1$3G for smokers compared to nonsmokers. A white woman of average age who doesn't walk much would have an average follicular phase $E_1$3G of 15 ng/mL if she doesn't smoke, and 20 ng/mL if she does. Follicular $E_1$3G also increases with ${}_1NAP$ or ${}_2FLUO$ that is higher than the other fluorene metabolites and decreases with ${}_2NAP$ or ${}_3FLUO$ higher than the other fluorene metabolites.

The final model for menstrual cycle length did not include any OH-PAHs when all cycles were included in the models. When only cycles with length within the normative range (21 to 35 days) were analyzed, cycle length increased with ${}_1PYR$ and ${}^d_2PHEN$ and decreased with ${}^d_9FLUO$.

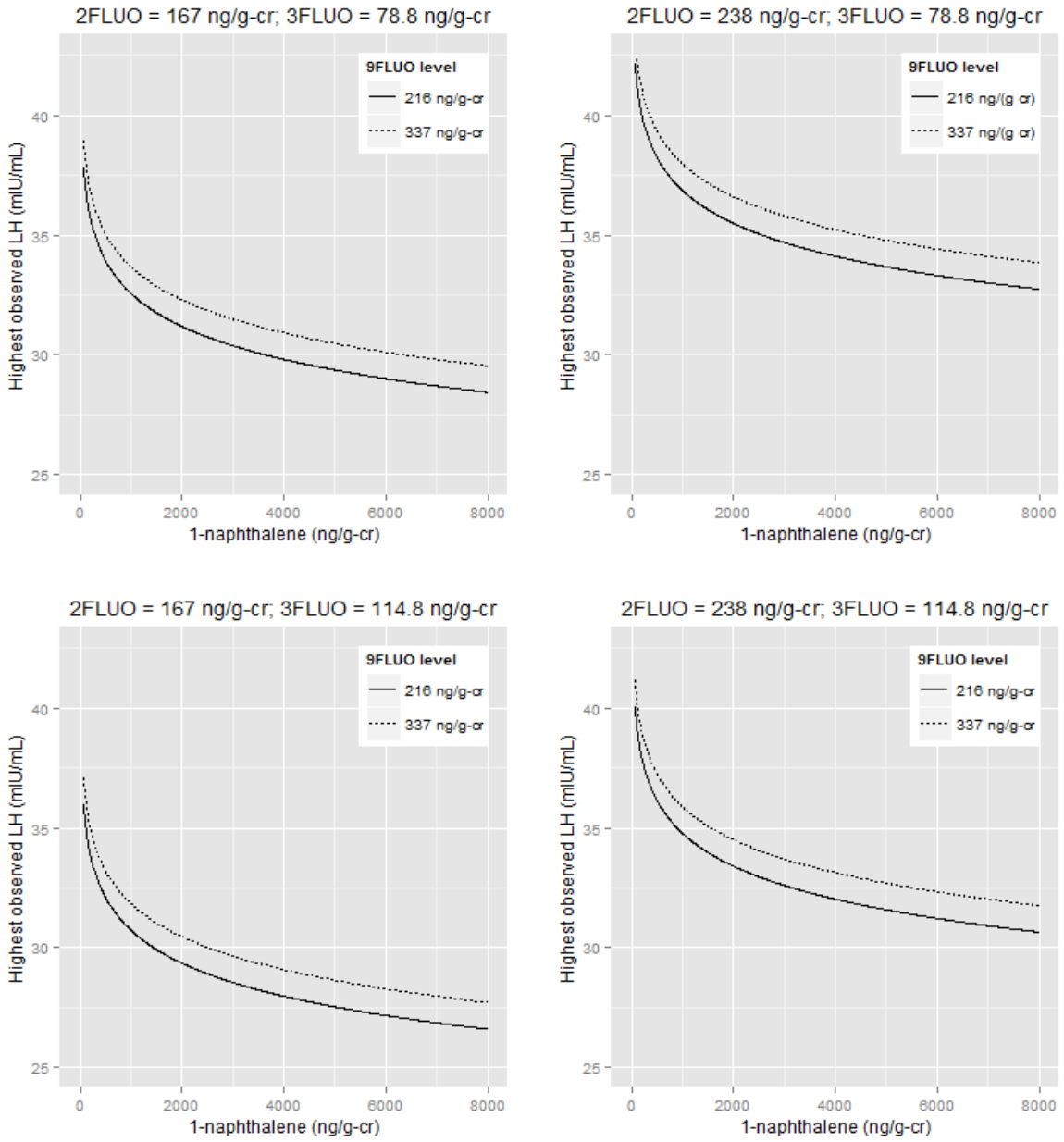

Figure 2.4: Graphical representation of changes in highest LH concentration with changes in concentrations of naphthalene and fluorene metabolites derived from the final model. All four graphs show the changes in highest urinary LH concentration with increasing urinary concentration of ${}_1NAP$ at two different concentrations of ${}_9FLUO$, the study median concentration (solid line) and the 75 percentile concentration (dotted line). ${}_2FLUO$ was held constant at its study median concentration in the two graphs on the left and at its 75th percentile in the two graphs on the right. ${}_3FLUO$ was held constant at its study median in the two upper graphs and at its 75th percentile in the two lower graphs. The graphs show that highest LH decreases with increasing ${}_1NAP$. A shift in ${}_2FLUO$ or ${}_9FLUO$ from the median to the 75th percentile results in approximately a 5 and 1 mIU/mL increase in highest LH, respectively while a similar shift in ${}_3FLUO$ results in about a 3 mIU/mL decrease in highest LH.

2.1.3 Discussion

Animal studies have clearly demonstrated that several higher molecular weight PAHs are potent ovotoxicants that destroy ovarian follicles, causing ovarian failure (Borman et al. 2000; Mattison 1979). Cigarette smoking is associated with decreased fecundity and earlier menopause in women, which may be due to the dozens of PAHs found in tobacco smoke (Harlow and Signorello 2000)Mattison89. However, this is the first study that has directly examined the associations of specific biomarkers of PAH exposure with measures of ovarian function in women. We found overall positive associations of fluorene metabolites and negative associations of naphthalene metabolites with two measures of LH surge amplitude. We found positive associations of ${}_2NAP$ and ${}_1PYR$ and negative association of phenanthrene metabolites with average follicular LH. Follicular phase length was positively associated with ${}^{d}_{3}FLUO$ and ${}_1PYR$. Smoking was positively associated with follicular and periovulatory $E_1$3G concentrations and $E_1$3G slope, and PAH metabolites were retained in these models.

A few prior studies have examined the effects of cigarette smoking on similar endpoints. Smoking has been associated with shorter cycle and follicular phase lengths (Windham, Elkin, et al. 1999), increased early follicular phase urinary $E_1$3G, non-significantly decreased luteal phase urinary pregnanediol-3-glucuronide, and higher urinary FSH concentrations during the luteal to follicular phase transition (Windham, Mitchell, et al. 2005). Similarly, increased early follicular phase serum estradiol and follicular phase progesterone concentrations have been reported in smokers compared to non-smokers (Zumoff et al. 1990). The same study reported decreased follicular phase serum LH concentrations and no differences in peak preovulatory LH concentrations. In contrast, another study reported that exposure to environmental tobacco smoke was associated with decreased mid to late follicular phase urinary $E_1$3G concentrations during non-conception cycles only (Chen et al. 2005).

It is difficult to compare the associations between OH-PAH metabolites and endocrine endpoints in the present study with the reported associations between exposure to tobacco smoke and similar endocrine endpoints in prior studies. Active smoking is associated with significantly increased urinary excretion of all nine OH-PAHs measured in the present study, while exposure to environmental tobacco smoke is associated with lesser or no increase in excretion of these metabolites (Aquilina et al. 2010; Suwan-Ampai et al. 2009). Tobacco smoke contains many high molecular weight PAHs, which were not directly biomonitored in the present study because their metabolites are primarily excreted in the bile and are below the limit of detection in urine (Li et al. 2008). Tobacco smoke also contains thousands of other compounds (Shopland et al. 2001), and therefore observed associations with smoking may or may not be due to the PAH content of tobacco smoke.

Smoking was retained in the second stage of modeling along with specific PAH metabolites only for the three $E_1$3G endpoints in the present study. This indicates that the OH-PAH metabolites contribute additional information to the models above and beyond that contributed by smoking. We observe increased follicular $E_1$3G concentrations with smoking, which is consistent with increased early follicular phase $E_1$3G in smokers reported by Windham, Mitchell, et al. (2005). We also observed associations of follicular $E_1$3G with fluorene, naphthalene, and phenanthrene metabolite concentrations. In contrast, our finding that periovulatory $E_1$3G increases with smoking is opposite the findings of Chen et al. regarding the association between environmental tobacco smoke exposure and mid to late follicular phase $E_1$3G in non-conception cycles (Chen et al. 2005). We also observed that periovulatory $E_1$3G decreases with ${}_2NAP$ and ${}_1PYR$ and increases with fluorene. While we find that peak and highest LH decreased with naphthalene and phenanthrene and increased with fluorene, we see no effect for smoking on these endpoints, concordant with prior findings of no effect of smoking on LH surge amplitude (Zumoff et al. 1990). While we find that both pyrene and the fluorene metabolite profile associate with follicular phase length, we see no

effect for smoking, which contrasts with prior findings associating shorter follicular phase length with smoking (Windham, Elkin, et al. 1999).

For women who do not smoke, food and exposure to particulate matter air pollution are the major sources of PAH exposures. Consumption of grilled meat is associated with greater increases in urinary $_1NAP$ and $_2NAP$ than the other PAH metabolites in the NHANES data (Suwan-Ampai et al. 2009). The OH-PAHs monitored in the present study have also been measured in relation to exposure to wood smoke. Compared to other studies that examined the effects of smoking on reproductive endpoints in women, the measurement of biomarkers that integrate PAH exposures from all routes is a strength of our study, and our models demonstrate that individual OH-PAH metabolites may have divergent associations with the same endpoint.

While the different associations we observed between various OH-PAH metabolites and several endocrine endpoints are interesting, we must be cautious in interpreting these associations because most of the OH-PAHs were highly correlated with one another. Only $_2NAP$ was not highly correlated with the other OH-PAHs. Interestingly, $_2NAP$ was a significant predictor for five of the eight endocrine endpoints. For the other OH-PAHs retained in seven of the nine models, their combined effects may be considered to be representative of the combined effects of all of the highly correlated OH-PAHs. In addition, because we examined a large number of endpoints and PAH metabolites, some of the associations may be spurious and should be confirmed in larger studies. It is also possible that associations were missed because we were only able to measure biomarkers of PAH exposure once per cycle during the follicular phase, reflecting exposures during the preceding day to a few days. In the future, it would be ideal to conduct a similar study in which PAH metabolites are measured more frequently during each cycle.

Another limitation of the present study is that 29% of the cycles were indeterminate for ovulatory status due to missed testing days. As a result, we had reduced power to detect

effects of OH-PAHs on ovulatory status. Missed test days also decreased the number of cycles for which some of the other endpoints could be calculated. A larger study will be required to confirm the associations between OH-PAHs and endocrine endpoints we observed.

In summary, we have demonstrated the feasibility of using urinary reproductive hormone data obtained via microelectronic fertility monitors to calculate endocrine endpoints for epidemiological studies of ovarian function during multiple menstrual cycles. We observed associations between biomarkers of environmental PAH exposure and follicular phase length, follicular phase LH and $E_1$3G concentrations, preovulatory LH surge amplitude, and peri-ovulatory $E_1$3Gconcentration and slope. The results show that environmental exposure to PAHs is associated with endocrine markers of ovarian function in women.

2.2 Shared Parameter Modeling

In this section, we discuss in detail the method we developed to deal with PAH data missingness in the study above, which we call shared parameter (SP) modeling. We begin by describing our motivation—why we believed our data necessitated this approach—and some alternative approaches to dealing with data similar to ours. We then give a detailed explanation of the method, along with some simulation results to show how the method behaves relative to two alternative approaches.

2.2.1 Background

A common problem faced by medical researchers is the expense of collecting certain types of data, such as chemical analysis data or brain imaging data. When funding is limited and these data are critical to a research question being pursued, researchers are often faced with difficult choices: Should we compromise statistical power by only looking at a small number

of expensive observational units? Should we compromise our ability to adequately address research questions by only collecting data on a cheaper subset of the variables we consider scientifically interesting?

In our study on the effect of environmental pollutants on the human menstrual cycle, our ability to obtain data on environmental pollutants (specifically OH-PAHs) was limited by grant funding. The necessary chemical testing of monthly urine samples to obtain OH-PAH concentration data is expensive, and although urine samples were collected for this purpose from each woman during each month of the study, our funding only permitted us to test half of our samples.

The SP modeling approach we developed for use with these data is an attempt to make the best use of our available data. We do not wish to throw out data corresponding to missing PAH values—there is still useful information in these data about our endpoints, as well as about the relationships between our baseline covariates and our endpoints. At the same time, we prefer not to impute PAH data for missing observations. The PAHs are our main covariates of interest, and we are missing half of the data on them. We are concerned that the imputation structure we choose for these data might have an outsized impact on our inferences.

The idea of SP modeling is based on a partitioning of a dataset into two subsets. The first subset, which we will call the *complete observations*, contains observations on every covariate of interest. The second subset, which we will call the *incomplete observations*, contains observations on only a subset of the covariate space. We assume that incomplete observations are all incomplete in the same way: they all include information on the same subset of covariates and they all lack information on the same subset of covariates. As an example, in the study we describe above, half of our observations are missing PAH information, but[2] these observations all have complete data on non-PAH covariates. We call the covariates for

[2]With some minor, easily imputed exceptions.

which values are recorded on each observational unit *complete covariates*; covariates with missing values in the incomplete observations will be called *incomplete covariates*. Complete observations have no missing covariates. Incomplete observations contain information only on complete covariates. We further assume that data for the *response* variables are complete in all observations.

Our approach builds two similar models—one for the complete observations and one for the incomplete observations—and constrains certain parameters shared by those two models to be equal. If we believe that these data all arise from the same generating processes and our missing covariate data are missing at random (MAR) or missing completely at random (MCAR), this is straightforward. For more about missing data, see Daniels and Hogan (2008) or Little and Rubin (2002).

2.2.1.1 Existing approaches

Before we describe our SP modeling approach in detail, we review other approaches for data like these.

Removing incomplete data

The easiest method for dealing with MCAR data—although this may be more problematic for MAR data—is to simply ignore any incomplete data. In the scenario we describe, this could take two forms: removing covariates where data are incomplete, or removing observations where data are incomplete.

Removing covariates where data are incomplete is an undesirable solution. If these covariates are assumed to be relevant predictors for some response, then not including them in models for that response will lead to less accurate predictions. This may seem like an appealing

strategy to some if data on a covariate are very sparse, but it is philosophically difficult to justify excluding known information that one believes to be relevant.

Removing observations where data are incomplete is also undesirable, but less problematic. The cost of removing incomplete observations is a loss of statistical power—there are fewer observations in such a dataset than in a dataset where these incomplete observations are retained. We are not using a reduced model when we believe a larger model to be better, as we do if we remove incomplete covariates—but we are nonetheless still excluding known information that we believe to be relevant.

Imputation

The next choice for dealing with MCAR data and the most common for MAR data is to impute the missing values based on existing information about the covariates that are missing. Imputation can be done in two ways, pointwise imputation and probabilistic imputation, which we detail below.

Pointwise imputation methods include methods like mean substitution and regression imputation. In mean substitution, which is designed for MCAR data, missing observations are replaced with the sample mean of all the other observations in the dataset on the same variable. In regression imputation, which is designed for MAR data where mean substitution would be inappropriate, a regression model for the missing data is constructed from the complete data and missing observations are replaced with predictions from this model. Pointwise imputation results in reduced variability for the missing data, since missing observations are always imputed from some deterministic model and do not include measures of uncertainty about the imputations. The SP modeling method presented below is based on pointwise imputation, but adds a variance component to the response model to account for this uncertainty.

Probabilistic imputation is often Bayesian, and in this case is easily implemented using MCMC methods. A probability distribution is assigned to missing values, and imputed values are probabilistic predictions (rather than deterministic ones) based on these models. That is, rather than imputing the missing data using a sample mean or a regression fitted value, at each iteration of an MC method, a value from the predictive distribution for that covariate is imputed. In this way, uncertainty about the imputation is automatically incorporated into the analysis. If a poor probability model is chosen for the missing data, inferences will likely suffer.

2.2.1.2 Motivation

In the environmental epidemiology study described above, we were confronted with a problem: our collaborators' primary research interest was whether PAHs affect various menstrual cycle endpoints, but PAH data were only available for half of their observations. How could we best make use of their data to address their research interests?

We disliked both of the obvious choices: drop the observations for which PAHs were unavailable, or probabilistically impute the missing PAH values[3]. As described in Section 2.1.1.4, our PAH data were missing completely at random, which would allow us to drop our incomplete observations; but dropping observations meant sacrificing half of our collaborators' data, and would have resulted in only 2-3 observations per woman. Having only 2-3 observations per woman would make it difficult to use random effects models for these data. We also found probabilistic imputation problematic. With half of our data on PAHs missing, and our data indicating that PAHs were approximately log-normal, we were not comfortable basing so much of our analysis depending on our decisions about how to model this missing data.

[3]Our strong preference is for probabilistic imputation over pointwise imputation.

Our goal is to reconcile complete observations with incomplete observations. First, we assumed that both models share the same structure for complete covariates (the baseline covariates). Next we modeled incomplete observations to estimate that model structure. And then we look at how the inclusion of our incomplete covariates (the OH-PAH concentrations) changes the model when they were present. Second, assuming that PAHs would be related to our endpoints, we inferred that error variances should be smaller among observations where the PAH data were available than among observations where these data were not available. If the PAHs are related to the endpoints, a model that includes them should make more accurate predictions than a reduced model without them.

Based on the first of these considerations, we recognized that we would need to center our PAH data. Models with and without the PAH data could not share the same parameter values for the baseline covariates unless (1) PAH data were centered, and (2) PAH data were orthogonal to baseline covariates. We did not orthogonalize the PAH data relative to the baseline covariates because correlations between these covariates were small, and because we wanted to preserve the original PAH variables for interpretability. Our simulations in Section 2.2.3 consider the performance of SP modeling when correlations exist between complete and incomplete covariates.

Based on the second of these considerations, we determined that an additional variance component should be added to the model for observations where PAH data were missing.

2.2.2 Shared Parameter Modeling

Notationally, let Y_C denote the vector of responses among the complete observations and Y_I denote the vector of responses corresponding to the incomplete observations. We will use the letters X and Z to denote the complete and incomplete covariates respectively; with X_C referring to the matrix of complete covariates for complete observations, X_I referring to com-

plete covariates observed on incomplete observations, Z_C referring to incomplete covariates observed on complete observations, and Z_I incomplete covariates on incomplete observations. Note that in our treatment, all information in X_C, X_I, and Z_C is known. Z_I is unknown.

In the linear model setting, we can now give our shared-parameter model a precise notation:

$$\begin{bmatrix} Y^C \\ Y^I \end{bmatrix} = \begin{bmatrix} X^C & Z^C \\ X^I & 0 \end{bmatrix} \begin{bmatrix} \beta \\ \gamma \end{bmatrix} + \begin{bmatrix} e^C \\ e^C + e^I \end{bmatrix}, \tag{2.1}$$

where β and γ are regression coefficients, and e^C and e^I are errors. The models for Y^C and Y^I can be thought of as a full model and a reduced model, respectively. Then e^C can be thought of as the error variance when all the covariates are available, and e^I as the additional error variance when there are missing covariates. Further, if $\text{Var}[e^I] = 0$, this means that the incomplete covariates are not improving the model fit for the complete data.

To discuss the underlying assumptions of shared-parameter modeling, it is useful to consider both the linear model formulation (Equation 2.1) as well as the split model formulation below:

$$\begin{aligned} Y^C &= X^C\beta + Z^C\gamma + e^C, & e^C &\sim N(0, V^C) \\ Y^I &= X^I\beta + e^C + e^I, & e^I &\sim N(0, V^I) \end{aligned} \tag{2.2}$$

From Equation (2.2), it is easy to see that $X\beta$ must play an identical role in both models. This requires that the covariates in X and Z be linearly independent, otherwise collinearity among predictors will cause the β coefficients to differ between the two models if they are fit separately. We examine the method's sensitivity to violations of the independent covariates assumption in Section 4.

Moreover, it is necessary that either the covariates in Z be standardized to have a zero mean, or Z include a column of ones corresponding to an intercept. Otherwise the value of an intercept coefficient in β will be different for the two models. No further study will be made of this assumption, because it should be straightforward for practitioners to address it in practice. Beyond these points, the general assumptions of linear modeling apply.

If the covariates in Z are assumed to have mean zero, we recognize that the model we describe is similar to a model using mean substitution to impute the values of the missing data. With mean substitution, if the covariates in Z have mean zero, then we add $0\gamma = 0$ to the incomplete data model when we impute. This is what we have done. The model for Y^C includes information on the incomplete covariates. The model for Y^I differs in two ways: (1) it does not include the incomplete covariates, which is equivalent to imputing a zero for each missing value; and (2) it includes an additional error term to quantify the difference in how well each model fits the response data. Point (2) provides a major distinction in the models and mitigates the issue of substituting zeroes for Z^I. We believe this is a sensible alternative, avoiding the assumption of a structured probability model on our missing data, 50% of which was unavailable to us.

This structure is particularly useful in the mixed modeling context where we have multiple observations within clusters, some of which include the incomplete covariates and some of which have missing values. Let $\mathbf{Y} = \{Y_i\} = \{y_{ij}\}$ be a vector of response data on clusters $i \in \{1, ..., k\}$, with $j \in \{1, ..., N_i^C, N_i^C + 1, ..., N_i^C + N_i^I\}$ observations per cluster. We let N_i^C be the number of complete observations in cluster i and N_i^I be the number of observations with

missing values in cluster i. Then for each cluster we can write

$$Y_i = \begin{bmatrix} Y_i^C \\ Y_i^I \end{bmatrix} = \begin{bmatrix} y_{i,1}^C \\ \vdots \\ y_{i,N_C}^C \\ y_{i,N_C+1}^I \\ \vdots \\ y_{i,N_C+N_I}^I \end{bmatrix}.$$

The random effects version of the SP model is

$$\begin{aligned} Y_i^C &= X_i^C\beta + Z_i^C\gamma + \eta_i J_{N_i^C} + e^C, \qquad & e^C &\sim N(0, V^C) \\ Y_i^I &= X_i^I\beta + \eta_i J_{N_i^I} + e^C + e^I, \qquad & e^I &\sim N(0, V^I) \end{aligned} \tag{2.3}$$

where J_d is a $d \times 1$ vector of ones and η_i is a cluster-specific random effect we will assume follows a $\mathrm{N}(0, \sigma_h^2)$.

SP modeling allows us to learn more about the random effects for each cluster from the incomplete observations on that cluster. In our environmental epidemiology dataset, this means we can learn more about individual women's response values by analyzing the data for which PAH values are missing. Since only 2-3 PAH values are available per woman, it would be difficult to accurately estimate the random effects for each woman without more information.

2.2.3 Simulation Results

To see how shared-parameter modeling can improve prediction accuracy when compared to fitting models on only complete observations/covariates, we examined results from a variety

of simulations. Data for the following simulations were generated from a linear mixed model with k clusters and N observations per cluster,

$$\begin{aligned} Y_{ij} &= X_{ij}\beta + Z_{ij}\gamma + \eta_i + e_{ij}, \\ \eta_i &\sim N(0, \tau^2), \\ e_{ij} &\sim N(0, \sigma^2), \end{aligned} \tag{2.4}$$

with $i \in \{1, ..., k\}$ and $j \in \{1, ..., N\}$. We will let X and Z be univariate here, so β and γ are scalars. Further, we set the residual variance for the full model, σ^2 equal to one in our simulations, and we set γ be equal to one in our simulations.

In each simulation, we consider the difference in predictive validity between the shared-parameter model, a simplified model using only the complete observations, and a model where all observations are considered but covariates with some missing values are removed. We are interested in how much information is to be gained from consideration of the incomplete observations. Our simulations consider the effect on predictive validity when we vary (1) the correlation between observations on the same subject, $\rho = \sqrt{\frac{\tau^2}{\tau^2+\sigma^2}}$; (2) the correlation between X and Z, ϕ; (3) the ratio of the number of complete observations per cluster, N^C, to the number of incomplete observations per cluster, N^I; (4) the total number of observations, n; (5) the number of subjects being studied, k; and (6) the strength of relationship between the complete covariate X and the response Y.

Results presented below are based on averaging over 100 simulations. In each simulation we randomly generate (i) kN independent values of X_{ij} and Z_{ij} from the appropriate bivariate normal distribution,

$$\begin{bmatrix} X_{ij} \\ Z_{ij} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \phi \\ \phi & 1 \end{bmatrix} \right);$$

(ii) values of the $k \times 1$ vector η, (iii) values of the $kN \times 1$ error vector e, and (iv) calculating the resulting $kN \times 1$ response vector Y. In our N observations per cluster, we include N^C complete

observations, N^I incomplete observations where we will treat the Z_{ij} data as missing, and $N^T = 3$ out-of-sample test observations. For each simulation, model fit was assessed via prediction error on these three test observations.

Our choice of three models is driven by what we see as relatively common choices that might be made by researchers facing a situation like we describe: where, because of expense or availability, data on some covariates are only available for a limited number of observations, and data on other covariates are available on those observations as well as others. We consider the shared-parameter model to be the most reasonable choice for this situation; but because of the additional complexity involved in using all of the data in this setting, we believe researchers may be likely to gravitate toward one of the other options presented. The first option, which we call "observations removed", fits a model based only on those observations for which complete covariate information is available, significantly decreasing the sample size available to the researcher. This option will lose power and information relative to the shared-parameter model, but is not unreasonable. The second option, which we call "covariates removed", ignores the covariates which only appear in a subset of the observations and models the data based only on the complete covariates. This alternative retains the full sample size and may be attractive from a power standpoint, but should be recognized as a poor option any time researchers believe the incomplete covariates to be effective predictors.

Table 2.8 presents results for simulations regarding the effects of elements (1) and (2) above. Note that prediction error under the SPM column and the observations removed column should go no lower than 1, the error variance built into the simulation. Prediction error under the covariates removed condition should be larger because it fails to account for the variance in Y attributable to $Z\gamma$. As ρ—the correlation between observations on the same subject—increases, the predictive performance of shared-parameter modeling improves relative to the two alternatives discussed. Even at $\rho = 0.0$, however, shared-parameter modeling

ρ	ϕ	Shared-Parameter Model	Observations Removed	Covariates Removed
0.0	0.0	1.015	1.022	2.031
	0.3	1.046	1.040	1.925
	0.7	1.112	1.022	1.543
0.3	0.0	1.162	1.218	2.167
	0.3	1.178	1.216	2.102
	0.7	1.235	1.218	1.657
0.7	0.0	1.217	1.331	2.262
	0.3	1.222	1.309	2.180
	0.7	1.280	1.312	1.733

Table 2.8: Mean squared prediction errors in 100 simulations from the model in Equation 2.4 with $k = 50$ and $N^C = N^I = 3$. True parameter values $\beta = \gamma = 1$ were used in these simulations.

outperforms the alternatives. Conversely, as the correlation between complete and incomplete covariates (ϕ) increases, the predictive performance of shared-parameter modeling loses ground to the observations removed alternative, although shared-parameter modeling still performs comparably well or better whenever the covariate correlation is small. Results regarding ϕ reflect the sensitivity of the shared-parameter modeling method to violations of the assumption that the covariates be independent, as discussed in Section 2. Note also here that the covariates removed alternative performs best when ϕ is large—that is when information in the incomplete covariates is partially captured by knowing the complete covariates.

N^C	N^I	Shared-Parameter Model	Observations Removed	Covariates Removed
3	3	1.217	1.331	2.262
	10	1.145	1.339	2.171
	30	1.072	1.337	2.051
5	3	1.167	1.213	2.291
	10	1.111	1.199	2.146
	30	1.052	1.194	2.049
10	3	1.096	1.110	2.179
	10	1.067	1.094	2.079
	30	1.005	1.059	2.035

Table 2.9: Mean squared prediction errors in 100 simulations from the model in Equation 2.4 with $\rho = 0.7$ and $\phi = 0.0$. True parameter values $\beta = \gamma = 1$ were used in these simulations.

Table 2.9 presents results for simulations regarding the effects of elements (3) and (4) above, the proportion of observations which are complete and the total number of observations. As previously stated, the shared-parameter modeling method is most useful for longitudinal data and assumes that complete and incomplete covariates are uncorrelated, so simulation results here and later assume $\rho = 0.7$ and $\phi = 0.0$. What we see here is that the predictive performance of the shared-parameter model *relative to* the observations removed model improves as N^I increases and diminishes as N^C increases. Knowing that the strength of within-subjects correlation is related to the predictive gain of shared-parameter modeling, the results here are intuitive. With large N^I, shared-parameter modeling has more observations with which to estimate a random effect. With large N^C, estimates of a random effect in the observations removed model will already be more accurate; the additional incomplete observations will do less to improve the precision of those estimates in the shared-parameter model.

k	β	Shared-Parameter Model	Observations Removed	Covariates Removed
5	0.2	1.423	1.602	2.421
	1.0	1.424	1.600	2.421
	5.0	1.369	1.548	2.446
15	0.2	1.243	1.370	2.390
	1.0	1.296	1.439	2.364
	5.0	1.269	1.404	2.217
50	0.2	1.202	1.300	2.308
	1.0	1.217	1.331	2.262
	5.0	1.233	1.332	2.344

Table 2.10: Mean squared prediction errors in 100 simulations from the model in Equation 2.4 with $\rho = 0.7$ and $\phi = 0.0$. In these simulations, $N^C = N^I = 3$ and $\gamma = 1$.

Table 2.10 presents results for simulations regarding the effects of elements (5) and (6) above—changes in the number of subjects observed, k, and the strength of the relationship between Y and X, summarized by β. Note that the strength of the relationship between Y and Z is held constant in these simulations; and since $\phi = 0$, the proportion of variability

in Y attributable to Z remains equal to the proportion of variability in Y attributable to unmodelable error ($\sigma^2 = 1$). We chose to vary the true value of β because the largest improvements in coefficient estimation under SP modeling occur for coefficients on complete covariates. We see little predictive improvement in SP modeling from varying β, however—a sign that the amount of response variability attributable to the complete covariate does not seem to significantly effect performance. Varying k appears to show small gains in predictive accuracy, suggesting that having more information to estimate the regression coefficients can provide some improvement beyond what is available from modeling the correlation structure in the data.

2.2.4 Future work

The work on SP modeling presented here has proven useful in our applied problem described in Section 2.1. Elaborating on this work by adding more mathematical rigor, providing clearer definitions, and more clearly explaining the interplay between the many simulation elements described above will allow us to develop this method for a more general audience. Moreover, our simulations have compared our method to naive removal of incomplete observations. It remains for us to compare our method's accuracy to different imputation methods and to two-stage regression methods.

Chapter 3

Marginalization for DIC – Part I

After beginning with a discussion of statistical model selection, this chapter will present technical details regarding the deviance information criterion (DIC) and explore its behavior in the mixed modeling setting. We discuss the mathematical and philosophical differences between using marginalized vs. unmarginalized DIC computations, and we offer two schemes for numerical approximation of the DIC in the linear mixed model (LMM) setting.

3.1 Background

In this section, we provide an introduction to the topic of statistical model selection and we review a number of important developments therein. We focus primarily on the class of model selection criteria known as information criteria, based on their connection to information theory and their interpretation as the information lost through modeling—a notion that arises from Kullback and Leibler (1951).

3.1.1 Philosophy of Model Selection

George Box famously said, "All models are wrong, but some are useful" (Box and Draper 1987). While this is a valuable dictum, in the area of model selection we must recognize that stochastic objects such as experimental data must, necessarily, arise from some stochastic process. Even if that process is unknowable, the fundamental goal of model selection is to identify—subject to certain constraints—which class of models best matches the unknown data-generating process.

Conceptually, we can consider that somewhere in the space of all probability distributions there exists a unique distribution by which our data were generated. If we consider a restricted subspace of probability distributions such as a parametric family, model selection seeks to find some distribution within this subset that comes as close as possible to replicating the behavior of the data-generating distribution. Following the conventions of D. J. Spiegelhalter et al. (2002), hereafter SBCV, we refer to the former generating distribution as the true distribution and the latter approximating distribution as a pseudo-true distribution. For a given true distribution there may be many different pseudo-true distributions, each corresponding to a different subspace of probability distributions. A pseudo-true distribution is often a parametric distribution, and in this case we refer to the parameter of a pseudo-true distribution as a pseudo-true parameter. For example, a set of data might be modeled by either a Weibull or a log-normal distribution. A pseudo-true Weibull distribution and a pseudo-true log-normal distribution would both exist for that data, each with their own pseudo-true parameters. These distributions would be the closest fit in each class to the true data-generating distribution, but neither would necessarily be that true distribution.

The fundamental goal of statistical model selection is to identify a model with good predictive accuracy for some set of response data y. Model selection tools often marry a measure of goodness-of-fit to a measure of desirability. For example, Akaike's information criterion

(Akaike 1974) uses cross entropy as a measure of goodness-of-fit and adds a complexity penalty equal to the number of parameters in the model. In this way, the criterion tends to pick models with fewer parameters when they yield comparable cross entropies, but if additional complexity can result in appreciably better goodness-of-fit, a larger model may be preferred. Note, however, that when models are picked purely through comparison of criteria such as this, one cannot guarantee that the resulting model fits the data well—only that it fits the data better than the other models considered.

Before we turn to our own work with the deviance information criterion (DIC), we review the basis for information criteria beginning with the Kullback-Leibler (KL) divergence and early information criteria. We develop and explain the ideas behind the DIC, as well as discuss two newer information criteria that have been created to address issues related to selection in hierarchical models. We do this to provide a fuller understanding of the framework surrounding DIC: both its historical place and how it relates to other criteria that are frequently used in model selection.

3.1.2 Kullback-Leibler (KL) Divergence, 1951

Of particular interest here is what it means for two distributions to be similar to one another. Traditionally, one would use a distance metric to express this. A common method has been to use the Kullback-Leibler (KL) divergence (Kullback and Leibler 1951).

Formally, consider two probability measures μ_0 and μ_1, both absolutely continuous with respect to a third measure λ. By the Radon-Nikodym theorem, for $j \in \{0,1\}$ there exist measurable functions f_j such that for a measurable set E under λ,

$$\mu_j(E) = \int_E f_j(y) d\lambda(y).$$

Kullback and Leibler define $\log \frac{f_0(y)}{f_1(y)}$ to be the information in y for discriminating between the hypothesis that y was selected from a population with probability measure μ_0 and the hypothesis that y was selected from a population with probability measure μ_1. The μ_0-directed KL divergence is defined as the expected amount of information for discriminating between these hypotheses contained in an observation from μ_0, namely

$$KL_{(\mu_0:\mu_1)} = \int \log\left(\frac{f_0(y)}{f_1(y)}\right) f_0(y) d\lambda(y).$$

The directed KL divergence is not a distance metric—it neither satisfies symmetry nor the triangle inequality. Even so, it is a premetric, satisfying the properties

$$KL_{(\mu_0:\mu_1)} \geq 0 \,\forall\, \mu_0, \mu_1 \qquad \text{and} \qquad KL_{(\mu_0:\mu_1)} = 0 \iff \mu_0 = \mu_1.$$

We can rewrite the formula for the directed divergence as

$$KL_{(\mu_0:\mu_1)} = \mathrm{E}_{\mu_0}[\log f_0(y)] - \mathrm{E}_{\mu_0}[\log f_1(y)]\,. \tag{3.1}$$

The term $-\mathrm{E}_{\mu_0}[\log f_0(y)]$ is known as the entropy of μ_0. The term $-\mathrm{E}_{\mu_0}[\log f_1(y)]$ is called the cross entropy of μ_0 and μ_1. Thus, the KL divergence is often conceptualized as the difference between the cross entropy of (μ_0, μ_1) and the entropy of μ_0 alone. The cross entropy forms the basis for many model selection procedures as we demonstrate below.

3.1.3 Akaike's Information Criterion (AIC), 1974

Akaike (1974) proposed an information-criterion-based model selection procedure, the AIC, that derives its validity from the interpretation of the negative cross entropy as a measure of the proximity between an inexact model and a true generating distribution, as demonstrated

with the KL divergence above. Akaike considers a scenario where μ_0 is the true generating distribution and μ_1 is an inexact modeling distribution whose fit we want to assess. He rewrites (3.1) as

$$KL_{(\mu_0:\mu_1)} = \int \log[f_0(y)] f_0(y) d\lambda y - \int \log[f_1(y)] f_0(y) d\lambda(y),$$

additively separating the component involving the modeling distribution from the component that involves only the generating distribution. Akaike shows how the above form can be used to compare different models, by recognizing that they share the unknown constant

$$c(\mu_0) = \int \log[f_0(y)] f_0(y) d\lambda y$$

that only involves the data-generating distribution.

What we are left with is an expectation over the data-generating distribution. Consider a set of alternate modeling distributions, $\mu_1, ..., \mu_M$. Then the directed KL divergence for μ_j relative to the generating distribution is given by

$$KL_{(\mu_0:\mu_j)} = c(\mu_0) - \mathrm{E}_{\mu_0}\big[\log f_j(y)\big] \qquad\qquad j \in \{1, ..., M\}.$$

Approximating this expectation with a sample mean of values that are generated by the unknown distribution μ_0 allows us to use the negative cross entropy as a measure of fit that can be compared across a range of modeling distributions.

Now assume our modeling distributions $\mu_1, ..., \mu_M$ are parametric, with parameter vectors $\theta_1, ..., \theta_M$ respectively. Assume further that the true distribution μ_0 is parametric, with θ_0. Then let a sample of data $y_i \overset{\text{iid}}{\sim} f(y_i \mid \theta_0)$, $i \in \{1, ..., n\}$ and a consider μ_j with density $f_j(y \mid \theta_j)$, $\theta_j \in \Theta_j$. There exists a maximum likelihood estimate (MLE) for θ_j based on these data, $\hat{\theta}_j$.

As Cavanaugh (1997) explains,

$$-2\sum_{i=1}^{n}\log f_j(y_i \mid \hat{\theta}_j)$$

is a biased estimator for twice the negative directed KL divergence between the true and fitted models,

$$-2\,\mathrm{E}_{\theta_0}\left[KL_{(\theta_0:\hat{\theta_j})}\right] = -2c(\mu_0) + 2\,\mathrm{E}_{\theta_0}\left[\sum_{i=1}^{n}\log f_j(y_i \mid \hat{\theta}_j)\right].$$

Further, this bias is asymptotically equal to twice the dimension of $\hat{\theta}_j$.

Then for this collection of models, Akaike writes

$$AIC_j = -2\sum_{i=1}^{n}\log f_j(y_i \mid \hat{\theta_j}) + 2k_j, \tag{3.2}$$

where k_j is the dimension of θ_j—the number of parameters θ_j includes. He is able to ignore $c(\mu_0)$ because it constitutes a fixed adjustment to each model, and is thus not useful in making comparing among those models.

Equation (3.2) marries a goodness-of-fit measure, the sample average cross entropy between θ_0 and θ_j, to a measure of model complexity, $2k_j$. This marriage is standard for information criteria, and the complexity penalty guards against overfitting. Because the MLE $\hat{\theta}_j$ is a function of the observed data, models with more parameters will tend to fit better than submodels that include only a subset of those parameters. Together, these measures converge to the cross entropy as long as θ_j is sufficiently close to θ_0. Model selection is performed by comparing AIC_j among a collection of models and choosing the model with the smallest AIC_j.

3.1.4 Bayes Factors (BF)

A common approach to model selection in the Bayesian setting is the Bayes factor (BF), the ratio of marginal likelihoods for the data under two distinct models. It is most easily understood in the context of hypothesis testing, where one of two models μ_1 or μ_2 is assumed to be the true distribution for some observed data, y. Our explanation below is derived from R. R. Christensen, Johnson, et al. (2010).

We define θ_1, θ_2 to be the parameters associated with μ_1, μ_2, and $f_1(y \mid \theta_1), f_2(y \mid \theta_2)$ their associated pdfs. In the Bayesian setting, we are also concerned with prior distributions on these parameters, $P_1(\theta_1 \mid \mu_1), P_2(\theta_2 \mid \mu_2)$; and prior probabilities for each model, q_1, q_2 where $q_1 + q_2 = 1$. We will use μ_T to denote the true model, whichever one it is.

The Bayes factor is based on the marginal predictive density for y,

$$f_j(y) = \int f_j(y \mid \theta_j) P_j(\theta_j \mid \mu_j) d\theta_j \qquad j \in \{1, 2\},$$

and the associated marginal likelihood $L(\mu_i \mid y) \propto f_j(y)$. The posterior probability of $\mu_1 = \mu_T$ is

$$\Pr[\mu_1 = \mu_T \mid y] = \frac{q_1 f_1(y)}{q_1 f_1(y) + q_2 f_2(y)}.$$

Then the posterior odds for $\mu_1 = \mu_T$ are

$$\begin{aligned} \frac{\Pr[\mu_1 = \mu_T \mid y]}{\Pr[\mu_2 = \mu_T \mid y]} &= \frac{\frac{q_1 f_1(y)}{q_1 f_1(y) + q_2 f_2(y)}}{\frac{q_2 f_2(y)}{q_1 f_1(y) + q_2 f_2(y)}} \\ &= \frac{q_1 f_1(y)}{q_2 f_2(y)} \\ &\equiv \frac{q_1}{q_2} BF \end{aligned}$$

Thus the Bayes factor comparing μ_1 to μ_2 is defined as

$$BF_{1:2} = \frac{f_1(y)}{f_2(y)}. \tag{3.3}$$

Since $\frac{q_1}{q_2}$ is the prior odds for $\mu_1 = \mu_T$, we can understand the Bayes factor as the degree to which our data change our prior beliefs about the odds. A Bayes factor above one means that the data favor the conclusion that $\mu_1 = \mu_T$, while a Bayes factor below one means that the data favor $\mu_2 = \mu_T$.

3.1.5 Bayesian Information Criterion (BIC), 1978

Schwarz (1978) provides the next major advance in the development of statistical information criteria. He begins by giving a concise summary of the criterion proposed by Akaike—quoted below[1]:

> An extension of the maximum likelihood principle is suggested... for the slightly more general problem of choosing among different models with different numbers of parameters. His suggestion amounts to maximizing the likelihood function separately for each model j, obtaining, say, $f_j(y_1, ..., y_n \mid \hat{\theta}_j)$, and then choosing the model for which $\log f_j(y_1, ..., y_n \mid \hat{\theta}_j) - k_j$ is largest, where k_j is the dimension of the model.

In contrast to this, Schwarz proposes that a model should instead minimize

$$BIC_j = -2 \sum_{i=1}^{n} \log f_j(y_i \mid \hat{\theta}_j) + k_j \log n, \tag{3.4}$$

[1]It is our standard practice in this dissertation, when quoting from other sources, to match their statements to our notation for the ease of the reader. We have endeavored to render the quoted material here and elsewhere as accurately as possible.

a value that what would later come to be called the Bayesian Information Criterion. Schwarz reasons asymptotically from the Bayesian strategy of picking the *a posteriori* most probable model from a class of models that are all given positive probability. This is very reminiscent of Equation (3.2). The only difference is the change in penalty term from $2k_j$ for AIC to $k_j \log n$ for BIC. The BIC penalty scales with the number of observed data values; and as more data are observed, BIC more strongly prefers a parsimonious model.

Formally Schwarz considers exponential family data with density $h(y)\exp\left(\theta t(y) - b(\theta)\right)$ where $\theta \in \Theta$ is multidimensional. Modeling distributions for these data, $\mu_1, ..., \mu_M$ depend on parameters $\theta_1, ..., \theta_M$ where θ_j lives in a k_j-dimensional subspace of Θ. Again, let q_j be the prior probability that model μ_j is correct, and let $P_j(\theta_j \mid \mu_j)$ be the prior distribution for θ_j conditional on μ_j. Then the Bayesian choice should select j to maximize

$$\begin{aligned}
S(j) &= \log\left(q_j f_j(y_1, ..., y_n)\right) \\
&= \log \int q_j \left(\prod_{i=1}^{n} h(y_i)\right) \exp\left(\theta_j \sum_{i=1}^{n} t(y_i) - nb(\theta_j)\right) dP_j(\theta_j \mid \mu_j) \\
&= \log\left(q_j \prod_{i=1}^{n} h(y_i) \int \exp\left(\theta_j \sum_{i=1}^{n} t(y_i) - nb(\theta_j)\right) dP_j(\theta_j \mid \mu_j)\right) \\
&= \log\left(\int \exp\left(\theta_j \sum_{i=1}^{n} t(y_i) - nb(\theta_j)\right) dP_j(\theta_j \mid \mu_j)\right) + \log q_j + \sum_{i=1}^{n} \log h(y_i)
\end{aligned}$$

Schwarz shows that

$$S(j) = \left(\hat{\theta}_j \sum_{i=1}^{n} t(y_i) - nb(\hat{\theta}_j)\right) - \frac{1}{2} k_j \log n + R_j,$$

where R_j is a remainder term bounded in n. As n grows, the boundedness of R_j means it goes away relative to $S(j)$ as a whole. This gives an asymptotic justification for using $k_j \log n$ as a complexity penalty in Equation (3.4) rather than Akaike's $2k_j$ in Equation (3.2). Of note is the fact that Schwarz's BIC relies on an asymptotic justification to eliminate prior beliefs about model preference—the prior model probability q_j is absorbed into the remainder R_j that is asymptotically eliminated.

Although the argument in Schwarz (1978) assumes the data come from an exponential family distribution, Cavanaugh and Neath (1999) show this result more generally. They let $Y = \{y_1, ..., y_n\}$ and let $f(Y)$ be the marginal density for the data over all models $\mu_1, ..., \mu_M$,

$$f(Y) = \sum_{l=1}^{M} q_l f_l(y_1, ..., y_n \mid \mu_l).$$

Then Cavanaugh and Neath show that

$$\begin{aligned} \log \Pr\left[\mu_j = \mu_T \mid Y\right] + \log f(Y) &\simeq \sum_{i=1}^{n} \log f_j(y_i \mid \hat{\theta}_j) - \frac{1}{2} k_j \log n \\ &= -\frac{1}{2} BIC_j \end{aligned}$$

Since $\log f(Y)$ does not depend on j, choosing a model based on $BIC(\theta_j)$ is asymptotically the same as choosing a model based on the posterior model probability.

From this, we can also see another interesting feature of the BIC: its asymptotic relation to the Bayes factor and the posterior odds. Observe that

$$\begin{aligned} \log \Pr\left[\mu_j = \mu_T \mid Y\right] + \log f(Y) &= \log\left(\frac{q_j f_j(Y \mid \mu_j)}{f(Y)}\right) + \log f(Y) \\ &= \log q_j + \log f_j(Y \mid \mu_j) \end{aligned}$$

Then using our definition of the BIC, (3.4), and the result from Cavanaugh and Neath (1999), we have

$$\begin{aligned} \log BF_{j:j'} &= \log f_j(Y \mid \mu_j) - \log f_{j'}(Y \mid \mu_{j'}) \\ &= \left(\log f_j(Y \mid \mu_j) + \log q_j - \log q_j\right) - \left(\log f_{j'}(Y \mid \mu_{j'}) + \log q_{j'} - \log q_{j'}\right) \\ &\simeq -\frac{1}{2}(BIC_j - BIC_{j'}) + (\log q_{j'} - \log q_j) \end{aligned}$$

If we assume equal prior probabilities for μ_j and $\mu_{j'}$, then, $\log BF_{j:j'} \simeq -\frac{1}{2}(BIC_j - BIC_{j'})$.

3.1.6 Log Pseudo-Marginal Likelihood (LPML), 1979

Geisser and Eddy (1979) argue that a fixed-penalty decision approach for choosing the wrong model, as Schwarz uses in his justification of the BIC, "may be reasonable for a selection procedure, but if the ultimate goal is prediction, then the penalty should depend both on sample size and type of error made." They build a predictive criterion, the log pseudo-marginal likelihood (LPML), using conditional predictive densities where each datapoint is fit knowing the rest of the data vector.

Let $Y_{(i)} = \{y_1, ..., y_{i-1}, y_{i+1}, ..., y_n\}$. This is the collection of all the elements in Y except y_i. Then for a model μ_j, Geisser and Eddy define a conditional predictive ordinate

$$CPO_{ij} = f_j(y_i \mid Y_{(i)}, \mu_j)$$

and a pseudomarginal likelihood

$$\begin{aligned} L_j &= \prod_{i=1}^{n} CPO_{ij} \\ &= \prod_{i=1}^{n} f_j(y_i \mid Y_{(i)}, \mu_j) \end{aligned}$$

A. Gelfand and Dey (1994) provide an easy method for calculating the inverses of the conditional predictive ordinates from a posterior sample $\theta_j^{(1)}, ..., \theta_j^{(B)}$,

$$\begin{aligned} CPO_{ij}^{-1} &= \mathrm{E}_{\theta_j|Y}\left[\frac{1}{f(y_i \mid \theta_j, \mu_j)}\right] \\ &\doteq \frac{1}{B}\sum_{s=1}^{B} \frac{1}{f(y_i \mid \theta_j^{(s)}, \mu_j)} \end{aligned}$$

The log pseudo-marginal likelihood itself, as the name implies, is given by

$$\begin{aligned} LPML_j &= \log L_j \\ &= \sum_{i=1}^{n} \log CPO_{ij}. \end{aligned}$$

Although not created as an information criterion, the LPML shares many of the same model selection uses as the common information criteria when comparing two models. It also has a standard interpretation as a "pseudo Bayes factor" (R. R. Christensen, Johnson, et al. 2010), and thus shares a connection with BIC in that they both provide approximations to the same quantity. Watanabe (2010a) also proves the asymptotic equivalence of LPML and the widely applicable information criterion (WAIC) presented in Section 3.1.8.

3.1.7 Deviance Information Criterion (DIC), 2002

Whereas Akaike and Schwarz are concerned with the deviation between a known model (expressed through $\hat{\theta}$) and the truth, the approach of SBCV considers an average deviation from the truth for a possible model. As before, let $Y = \{y_1, ..., y_n\}$ be a set of observed data, and let $\mu_1, ..., \mu_M$ be a collection of models with associated parameters $\theta_1, ..., \theta_M$ and pdfs $f_1(Y \mid \theta_1), ..., f_M(Y \mid \theta_M)$. Instead of using the cross entropy as defined in Section 3.1.2, SBCV consider the posterior expectation of the log density. They define

$$DIC_j = -2\, \mathrm{E}_{\theta_j|Y,\mu_j}\left[\log f_j(Y \mid \theta_j)\right] + p_{Dj}, \tag{3.5}$$

where p_{Dj} is a penalization term that will be defined below.

SBCV frame their method around a quantity $D(\theta_j)$ defined as

$$D(\theta_j) = -2\left[\log f_j(Y \mid \theta_j) - \log f(Y)\right],$$

where SBCV describe $f(Y)$ as "some fully specified standardizing term that is a function of the data alone." Observe that $f(Y)$ will play a similar role to $c(\mu_0)$ in the AIC development above—as a constant term not based on the model being evaluated. Although it appears in the formal development of DIC, because it is an empirical function of the data alone, it will be irrelevant in comparing DICs for different models.

Based on $D(\theta_j)$, SBCV then define the quantities

$$\overline{D(\theta_j)} = -2\,\mathrm{E}_{\theta_j|Y,\mu_j}\left[\log f_j\left(Y \mid \theta_j\right)\right] + 2\log f(Y)$$
$$D(\hat{\theta}_j) = -2\log f_j\left(Y \mid \hat{\theta}_j\right) + 2\log f(Y).$$

where $\hat{\theta}_j$ is some posterior summary for θ_j—most commonly a posterior mean, median, or mode. SBCV describe $\overline{D(\theta_j)}$ as a "Bayesian measure of fit [or] perhaps better considered a measure of 'adequacy'." It it the posterior expectation of their $D(\theta_j)$, and it quantifies how well the model fits on average, across the posterior distribution for θ_j. Meanwhile, they describe $D(\hat{\theta}_j)$ as a "classical 'plug-in' measure of fit," akin to the quantities used in the AIC and BIC.

This classical measure will tend fit be better[2] than SBCV's Bayesian measure—and in fact this is guaranteed to be the case when $f_j(Y \mid \theta_j)$ is log-concave in θ_j and we choose $\hat{\theta}_j$ to be the posterior mean of θ_j as SBCV recommend. Then Jensen's inequality guarantees

$$\log f_j\left(Y \mid \mathrm{E}_{\theta_j|Y,\mu_j}\left[\theta_j\right]\right) \geq \mathrm{E}_{\theta_j|Y,\mu_j}\left[\log f_j\left(Y \mid \theta_j\right)\right],$$

which is the same as saying that the log-likelihood evaluated at the estimate $\hat{\theta}_j$ is larger than the posterior expectation of the log-likelihood function for θ_j.

[2]Because SBCV have structured their work around the idea of deviances, "better" can be difficult to follow here. $D(\hat{\theta}_j)$ is better than $\overline{D(\theta_j)}$ when $D(\hat{\theta}_j) < \overline{D(\theta_j)}$.

The penalization term, p_{Dj}, is then defined as

$$\begin{aligned} p_D &= \overline{D(\theta_j)} - D(\hat{\theta}_j) \\ &= -2\,\mathrm{E}_{\theta_j|Y,\mu_j}\left[\log f_j\left(Y \mid \theta_j\right)\right] + 2\log f_j\left(Y \mid \hat{\theta}_j\right) \\ &= 2\left[\log f_j\left(Y \mid \hat{\theta}_j\right) - \mathrm{E}_{\theta_j|Y,\mu_j}\left[\log f_j\left(Y \mid \theta_j\right)\right]\right], \end{aligned} \tag{3.6}$$

SBCV interpret this quantity as the degree of overfitting when a classical measure of fit, $D(\hat{\theta}_j)$, is used in place of a Bayesian measure of fit, $\overline{D(\theta_j)}$.

In models where the likelihood admits a normal approximation, SBCV argue that p_D is approximately the number of free parameters in the model. We give a more general argument below that also suggests an asymptotic equivalency between DIC_j and AIC_j under the conditions that $L_j(\theta_j \mid Y)$, the likelihood for θ_j, admit a normal approximation; and the prior $p_j(\theta_j \mid \mu_j)$ is sufficiently diffuse.

To begin, let $\hat{\theta}_j^{ML}$ be the MLE for θ_j and let $\hat{\theta}_j^{B}$ be the posterior mode for θ_j. We repeat Equations (3.2) and (3.5):

$$\begin{aligned} AIC_j &= -2\sum_{i=1}^{n}\log f_j\left(y_i \mid \hat{\theta}_j^{ML}\right) + 2k_j \\ DIC_j &= -2\,\mathrm{E}_{\theta_j|Y,\mu_j}\left[\log f_j(Y \mid \theta_j)\right] + p_{Dj} \end{aligned}$$

Now observe that since

$$p_{Dj} = -2\,\mathrm{E}_{\theta_j|Y,\mu_j}\left[\log f_j\left(Y \mid \theta_j\right)\right] + 2\log f_j\left(Y \mid \hat{\theta}_j^{B}\right),$$

we can also write

$$-2\,\mathrm{E}_{\theta_j|Y,\mu_j}\left[\log f_j\left(Y \mid \theta_j\right)\right] = p_{Dj} - 2\log f_j\left(Y \mid \hat{\theta}_j^{B}\right).$$

Then

$$DIC_j = -2\sum_{i=1}^{n} \log f_j\left(y_i \mid \hat{\theta}_j^B\right) + 2p_{Dj}.$$

To show $AIC_j \simeq DIC_j$, it is sufficient to show that $\hat{\theta}_j^{ML} \doteq \hat{\theta}_j^B$ and $k_j \doteq p_{Dj}$.

It is well known that when the likelihood admits a normal approximation and the prior is sufficiently diffuse; $\hat{\theta}_j^{ML}$, $\hat{\theta}_j^B$, and $\mathrm{E}_{\theta_j|Y,\mu_j}[\theta_j]$ are consistent estimators of the same quantity (Gelman et al. 2013, p.92), so asymptotic equivalence is clear. Further, the large sample approximation to the posterior is known to be

$$\theta_j \mid Y, \mu_j \dot{\sim} \mathrm{N}\left(\hat{\theta}_j^B, 2\ddot{D}(\hat{\theta}_j^B)^{-1}\right),$$

as discussed in Gelman et al. (2013, p.93).

What remains to be shown is that p_{Dj} is asymptotically equal to the number of parameters in the model, k_j. Using a second-order Taylor expansion, we observe that

$$D(\theta_j) \simeq D(\hat{\theta}_j^B) + \frac{1}{2}\left(\theta_j - \hat{\theta}_j^B\right)^T \ddot{D}(\hat{\theta}_j^B)\left(\theta_j - \hat{\theta}_j^B\right),$$

where the first-order Taylor term is zero because our choice of $\hat{\theta}_j^B$ ensures that $\dot{D}(\hat{\theta}_j^B) = 0$. Then as David Hinkley would argue,

$$\begin{aligned}
\overline{D(\theta_j)} &= \mathrm{E}_{\theta_j|Y,\mu_j}[D(\theta_j)] \\
&\simeq \mathrm{E}_{\theta_j|Y,\mu_j}\left[D(\hat{\theta}_j^B) + \frac{1}{2}\left(\theta_j - \hat{\theta}_j^B\right)^T \ddot{D}(\hat{\theta}_j^B)\left(\theta_j - \hat{\theta}_j^B\right)\right] \\
&= D(\hat{\theta}_j^B) + \frac{1}{2}\,\mathrm{E}_{\theta_j|Y,\mu_j}\left[\left(\theta_j - \hat{\theta}_j^B\right)^T \ddot{D}(\hat{\theta}_j^B)\left(\theta_j - \hat{\theta}_j^B\right)\right]
\end{aligned}$$

Then since $\overline{D(\theta_j)} = D(\hat{\theta}_j^B) + p_{Dj}$, this indicates that

$$p_{Dj} \simeq \frac{1}{2}\,\mathrm{E}_{\theta_j|Y,\mu_j}\left[\left(\theta_j - \hat{\theta}_j^B\right)^T \ddot{D}(\hat{\theta}_j^B)\left(\theta_j - \hat{\theta}_j^B\right)\right].$$

We now show that

$$\frac{1}{2}\,\mathrm{E}_{\theta_j|Y,\mu_j}\left[\left(\theta_j-\hat{\theta}_j^B\right)^T\ddot{D}(\hat{\theta}_j^B)\left(\theta_j-\hat{\theta}_j^B\right)\right]\doteq k_j.$$

Since $\mathrm{E}_{\theta_j|Y,\mu_j}\left[\left(\theta_j-\hat{\theta}_j^B\right)^T\ddot{D}(\hat{\theta}_j^B)\left(\theta_j-\hat{\theta}_j^B\right)\right]$ is a scalar, we can also write

$$\mathrm{E}_{\theta_j|Y,\mu_j}\left[\left(\theta_j-\hat{\theta}_j^B\right)^T\ddot{D}(\hat{\theta}_j^B)\left(\theta_j-\hat{\theta}_j^B\right)\right]=\mathrm{tr}\left(\ddot{D}(\hat{\theta}_j^B)\,\mathrm{E}_{\theta_j|Y,\mu_j}\left[\left(\theta_j-\hat{\theta}_j^B\right)\left(\theta_j-\hat{\theta}_j^B\right)^T\right]\right).$$

But our large sample posterior approximation gives

$$\begin{aligned}\mathrm{E}_{\theta_j|Y,\mu_j}\left[\left(\theta_j-\hat{\theta}_j^B\right)\left(\theta_j-\hat{\theta}_j^B\right)^T\right]&\doteq\mathrm{Cov}_{\theta_j|Y,\mu_j}\left[\theta_j\right]\\&\doteq 2\ddot{D}(\hat{\theta}_j^B)^{-1}.\end{aligned}$$

Then

$$\begin{aligned}p_{Dj}&\simeq\frac{1}{2}\,\mathrm{E}_{\theta_j|Y,\mu_j}\left[\left(\theta_j-\hat{\theta}_j^B\right)^T\ddot{D}(\hat{\theta}_j^B)\left(\theta_j-\hat{\theta}_j^B\right)\right]\\&\doteq\frac{1}{2}\,\mathrm{tr}\left(\ddot{D}(\hat{\theta}_j^B)\left(2\ddot{D}(\hat{\theta}_j^B)^{-1}\right)\right)\\&=\mathrm{tr}\left(\ddot{D}(\hat{\theta}_j^B)\ddot{D}(\hat{\theta}_j^B)^{-1}\right)\\&=k_j\end{aligned}$$

The interpretation of p_D as the number of free parameters in a model is widely considered to hold more generally than we argue above, but not without debate. The choice of a posterior summary $\hat{\theta}_j$ can affect p_{Dj}, even to the point of making it negative. There is also no guarantee that p_{Dj} will be positive when the posterior mean is chosen in cases where the density is not log-concave in θ_j, as is often the case with mixture models. Further, the DIC is not well defined in hierarchical models. Celeux et al. (2006) present eight possible DIC constructions for hierarchical models, differing in how the latent parameters are handled by

$\overline{D(\theta_j)}$ and $D(\hat{\theta}_j)$. We discuss these issues further in the following sections, and we present three DIC constructions pertaining to mixed modeling in Sections 3.3.1 below.

It is important to note also that DIC_j as it has been defined is not analytically tractable, although it is simple to numerically approximate it given an MCMC sample from the posterior distribution, θ_j, $\{\theta_j^{(1)}, ..., \theta_j^{(B)}\}$. Below is the standard computational form of DIC_j when $\hat{\theta}_j$ is taken to be the posterior mean.

$$DIC_j \simeq -\frac{4}{B}\sum_{s=1}^{B} \log f_j\left(Y \mid \theta_j^{(s)}\right) + 2\log f_j\left(Y \mid \frac{1}{B}\sum_{s=1}^{B}\theta_j^{(s)}\right)$$

3.1.8 Other Information Criteria

The Bayesian Predictive Information Criterion (BPIC; Ando 2007) and the Widely Applicable Information Criterion (WAIC; Watanabe 2010b) are two newer information criteria created to deal with the problems inherent in selection involving hierarchical models.

The first of these, the BPIC proposed by Ando in 2007, takes the form

$$\begin{aligned} BPIC_j &= -2\,\mathrm{E}_{\theta_j|Y,\mu_j}\left[\log f_j(Y \mid \theta_j)\right] + 2n\hat{b}_{\theta_j} \\ &= \left(\overline{D(\theta_j)} - 2\log f(Y)\right) + 2n\hat{b}_{\theta_j} \end{aligned}$$

where μ is a parametric model for y with parameter vector θ. Computation of the penalization term $\hat{b}_\theta$ is quite involved, but it is meant to approximate the bias created by using $\overline{D(\theta_j)}$ rather than $\mathrm{E}_{\mu_0}[D(\theta_j)]$ as a measure of model fit, where μ_0 is the true generating distribution for the data y as before.

Formally,

$$b_{\theta_j} = \int \left(\frac{1}{n}\,\mathrm{E}_{\theta_j|Y,\mu_j}\left[\log f_j(Y \mid \theta_j)\right] - \mathrm{E}_z\left[\mathrm{E}_{\theta_j|Y,\mu_j}\left[\log f_j(z \mid \theta_j)\right]\right]\right) d\mu_0(y),$$

where μ_0 is the true data-generating distribution, and where $z \sim \mu_0$. BPIC is, thus, using the same measure of fit—or as SBCV say, "adequacy"—as DIC, but with a different choice of penalization term to approximate the asymptotic bias in $\overline{D(\theta_j)}$ when the data are generated under an unknown distribution.

WAIC, proposed by Watanabe in 2010, is an attempt to build a Bayesian criterion that does not rely on plug-in point estimates of parameters. It depends on what Gelman et al. (2013) call the log pointwise predictive density (LPPD) for a model μ_j,

$$\begin{aligned} LPPD_j &= \sum_{i=1}^{n} \log f_j(y_i \mid Y, \mu_j) \\ &= \sum_{i=1}^{n} \log \int f_j(y_i \mid \theta_j) P_j(\theta_j \mid Y, \mu_j) d\theta_j \end{aligned}$$

where $P_j(\theta_j \mid Y, \mu_j)$ is the posterior density for θ_j. If a sample $\{\theta_j^{(1)}, ..., \theta_j^{(B)}\}$ is available from this posterior, then $LPPD_j$ can be numerically approximated with

$$LPPD_j \simeq \sum_{i=1}^{n} \log \left(\frac{1}{B} \sum_{s=1}^{B} f_j(y_i \mid \theta_j^{(s)}) \right).$$

Then $WAIC_j$ is defined as

$$WAIC_j = LPPD_j + p_{WAIC_j},$$

where p_{WAIC_j} is an overfitting penalty. Although Gelman et al. (2013, p.173) give two versions of this penalty, we do not intend to do an exhaustive review of the criterion here and will

only report the first.

$$
\begin{aligned}
p_{WAIC_j} &= 2\sum_{i=1}^{n}\Big(\log f_j(y_i \mid Y, \mu_j) - \mathrm{E}_{\theta_j|Y,\mu_j}\big[\log f_j(y_i \mid \theta_j)\big]\Big) \\
&= 2\sum_{i=1}^{n}\Big(\log \mathrm{E}_{\theta_j|Y,\mu_j}\big[f_j(y_i \mid \theta_j)\big] - \mathrm{E}_{\theta_j|Y,\mu_j}\big[\log f_j(y_i \mid \theta_j)\big]\Big) \\
&\simeq 2\sum_{i=1}^{n}\left(\log\left(\frac{1}{B}\sum_{s=1}^{B} f_j(y_i \mid \theta_j^{(s)})\right) - \frac{1}{B}\sum_{s=1}^{B}\log f_j(y_i \mid \theta_j^{(s)})\right)
\end{aligned}
$$

WAIC has a number of nice properties. Principally, it does not rely on point estimation as AIC, BIC, and DIC do. It has also been shown to be asymptotically equivalent to LPML (Watanabe 2010a), as well as to Bayesian leave-one-out cross-validation (Gelman et al. 2013, p.176).

3.2 Model Selection in Mixed Models

As we mentioned in Section 3.1.7, mixed modeling is an area where the DIC is not well defined and many competing constructions have been offered (Celeux et al. 2006). In this section we introduce the mixed modeling framework. We then discuss how the definition of p_D in particular is complicated by these models. We close the section with an example to demonstrate the behavior of p_D in a simple random effects model.

3.2.1 The Mixed Modeling Framework

Mixed models, models that incorporate both fixed and random effects, are commonly used in statistical analysis. To understand their appeal, consider a simple linear regression of a response y on a covariate x.

Let y and x both be measured at multiple times on multiple individuals. Denote as y_{ij} the response on individual i at time j, and define x_{ij} analogously. In a simple linear regression we may write $y_{ij} = \beta_0 + x_{ij}\beta_1 + e_{ij}$, where β_0 is the intercept of a regression line, β_1 is the slope of that line, and e_{ij} is the amount by which y_{ij} differs from the value that would be predicted for it based on the regression line. The standard assumption is that $e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$ for some constant variance σ^2 across all observations.

A linear mixed model (LMM) with random intercepts for each individual would be written as $y_{ij} = \beta_0 + x_{ij}\beta_1 + \gamma_i + e^*_{ij}$. Note that both of these models can apply to the same set of response and covariate data. In the LMM, we assume that $\gamma_i \overset{\text{iid}}{\sim} N(0, \tau^2)$ and $e^*_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2_r)$, with the γ's and e^*'s independent of each other. The total variability in the response data around the regression line is unchanged, but now we are splitting it into two terms: one that represents variability shared by observations on the same individual (γ_i) and another that represents leftover error that can't be attributed to individuals (e^*_{ij}). Since the total variability is the same, it is easy to see that $\sigma^2_r \leq \sigma^2$ and $\tau^2 \leq \sigma^2$. Because much of statistical inference depends on the amount of error in a dataset, using mixed models to account for between-subject variability allows statisticians to obtain more precise results when such variability exists. When such variability does not exist, $\tau^2 = 0$ and $\sigma^2_r = \sigma^2$, and the modeling cost incurred is simply that of estimating one extra parameter.

We proceed to give a mathematical definition for the mixed model that we use throughout the next three chapters.

Let $\mathbf{Y} = \{Y_i\} = \{y_{ij}\}$ be a $kn \times 1$ vector of response data on individuals $i \in \{1, ..., k\}$, with $j \in \{1, ..., n\}$ observations per individual. We use a balanced design with common n for all individuals to simplify some of the following linear algebra, but the results we obtain do not require this balance.

Let β be a $p \times 1$ vector of regression parameters. Let $\mathbf{X}$ be the $kn \times p$ design matrix for the regression parameters, $\mathbf{X}_i$ be the $n \times p$ block of the $\mathbf{X}$ matrix corresponding to cluster i, and X_{ij} be the $1 \times p$ row vector corresponding to the j^{th} observation on cluster i.

Let $\boldsymbol{\gamma} = \begin{bmatrix} \gamma_1^T & ... & \gamma_k^T \end{bmatrix}^T$ be a $kq \times 1$ vector of random effects, with γ_i the $q \times 1$ vector of random effects corresponding to cluster i. Let $\mathbf{Z}$ be the $kn \times kq$ block diagonal design matrix for the random effects. Let $\mathbf{Z}_i$ be the $n \times q$ submatrix of $\mathbf{Z}$ corresponding to its i^{th} diagonal block, and Z_{ij} be the $1 \times q$ row vector corresponding to the j^{th} row of the $\mathbf{Z}_i$ matrix.

Let $\boldsymbol{\psi} = \begin{bmatrix} \psi_1^T & ... & \psi_k^T \end{bmatrix}^T$ be the mean of the random effects vector $\boldsymbol{\gamma}$, and let $\boldsymbol{\Sigma}$ be block diagonal Σ_i, $i \in \{1, ..., k\}$ be the covariance matrix of the random effects. We assume that $\boldsymbol{\gamma} \sim \mathrm{N}_{kq}(\boldsymbol{\psi}, \boldsymbol{\Sigma})$, or equivalently here that $\gamma_i \overset{\text{indep}}{\sim} \mathrm{N}_q(\psi_i, \Sigma_i)$. We use θ to refer to the collection of parameters $\{\beta, \boldsymbol{\psi}, \boldsymbol{\Sigma}\}$.

Then the linear mixed model can be written as

$$\begin{aligned} &y_{ij} = X_{ij}\beta + Z_{ij}\gamma_i + e_{ij}, \\ &\gamma_i \overset{\text{indep}}{\sim} \mathrm{N}_q(\psi_i, \Sigma_i), \\ &e_{ij} \overset{\text{iid}}{\sim} \mathrm{N}\Big(0, \sigma^2\Big). \end{aligned} \tag{3.7}$$

Or equivalently

$$\begin{aligned} &\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e}, \\ &\boldsymbol{\gamma} \sim \mathrm{N}_{kq}(\boldsymbol{\psi}, \boldsymbol{\Sigma}), \\ &\mathbf{e} \sim \mathrm{N}_{kn}\Big(0_{kn}, \sigma^2 I_{kn}\Big). \end{aligned} \tag{3.8}$$

3.2.1.1 The role of y and x

Both y and x are observed values gathered by researchers. As Bayesians, we consider these values to be fixed—*statheric* nodes, from the Greek word for 'constant', as contrasted with

the *stochastic* (random) nodes θ and γ. In a mixed model, we always assume some sort of structure in the observations that allows us to account for part of the observed response variability by grouping "like" observations together. Usually this is indicative of some type of observational unit or cluster: individuals who have been observed repeatedly, hospitals where data were gathered on multiple patients, nations whose economic output is observed over a number of years.

3.2.1.2 The role of θ

We use θ here, and throughout the next two chapters, to refer to the collection of all parameters in a model. In the simple linear regression example described above, $\theta = \{\beta_0, \beta_1, \sigma^2\}$ before random effects are added and $\theta = \{\beta_0, \beta_1, \tau^2\sigma_r^2\}$ when they are included. In the general form, θ refers to the collection of parameters $\{\beta, \boldsymbol{\psi}, \boldsymbol{\Sigma}, \sigma^2\}$.

3.2.1.3 The role of γ

The random terms, γ, allow us to efficiently account for unexplained subject-level variability without having to add a fixed-effect parameter for each individual. When random effects, γ, are included in a model, this is tantamount to making a statement that there are individual-level differences in the response—baseline differences and/or differences in covariate effect on the response—that are not captured by the fixed covariate effects model. Random effects act as a catchall for structural elements that the statistician hasn't built into a model. They are expressed efficiently because the only parameter they add to the model is a variance term for the individual-level differences.

The elements of γ are not themselves parameters, but are more accurately thought of as latent random variables. They are unknown stochastic objects whose inclusion can help us better understand our response data. The linear mixed model can be written in such

a way that γ is never specified—see the section below on preprocessing marginalization. Generalized linear mixed models (GLMMs), where the response data are assumed to arise from some non-normal distribution, do not allow for this convenient marginalization; but the role of γ as a latent random vector is the same.

3.2.2 Complications with p_D in Mixed Models

In this section, we will explain why p_D is not well defined for mixed models and how this relates to SBCV's notion of a model "focus", the collection of stochastic objects one is interested in. We then give an example showing how p_D can differ considerably, even in a simple model, depending on the focus one chooses.

In their initial paper on the Deviance Information Criterion (DIC), SBCV identify a key concern in applying DIC to mixed- and hierarchical models:

> Since the complexity [penalty p_D] depends on the focus, a decision must be made whether nuisance parameters, e.g. variances, are to be included in [the collection of model parameters] Θ or integrated out before specifying the model $P(x \mid \theta, \mu)$. However, such a removal of nuisance parameters may create computational difficulties.

To prevent confusion with traditional statistical notion for parameters (that do not admit the "nuisance parameters" mentioned by SBCV), we again distinguish between stochastic and statheric objects in a Bayesian model. Data and statically defined parameters for priors in the model are statheric: fixed by the analyst and not subject to MCMC sampling. All other objects—parameters for the data distribution and latent variables—are stochastic. In the parlance of SBCV, stochastic objects include both focal parameters (objects that interest the researcher) and nuisance parameters (objects that do not).

The crux of this issue is embedded in the definition of θ in Equations (3.5) and (3.6). If θ is the collection of parameters in a model, then the *DIC* should provide a reasonable model selection criterion. However, if θ is defined more generally as the collection of all stochastic objects in a Bayesian model, this becomes problematic. We demonstrate this with the following example.

Consider a simple case discussed in Hodges and Sargent (2001), a traditional random effects model, and see how this issue manifests. Let $\mathbf{Y} = \{y_{ij} : i \in \{1, ..., k\}, j \in \{1, ..., n\}\}$, where y_{ij} represents the j-th measurement of some variable on an individual i, and let $N = k * n$. Here and throughout this work, we use I_n to refer to an n-dimensional identity matrix, J_n to refer to an $n \times 1$ vector of 1's, and J_n^n to refer to an $n \times n$ matrix of 1's. We write this random effects model as

$$\begin{aligned} y_{ij} &= \gamma_i + \varepsilon_{ij} \\ \gamma_i &\sim \mathrm{N}\left(\psi, \frac{1}{\tau_g}\right) \\ \varepsilon_{ij} &\sim \mathrm{N}\left(0, \frac{1}{\tau_e}\right), \end{aligned} \tag{3.9}$$

or equivalently

$$\mathbf{Y} \sim \mathrm{N}_N\left(\psi J_N, \left(\frac{1}{\tau_g} I_k \otimes J_n^n + \frac{1}{\tau_e} I_N\right)\right).$$

The parameters in this model are $\theta = \{\psi, \tau_g, \tau_e\}$. As previously discussed, the $k \times 1$ vector of $\boldsymbol{\gamma}$'s can be thought of as latent variables—unknown random objects that model correlation and extra heterogeneity in the data. By Equation (3.6), we can define a marginal p_D construction for this model:

$$\begin{aligned} p_{Dm} &= -2\left(\mathrm{E}_{\theta|y}[\log f(y \mid \theta)] - \log f\left(y \mid \mathrm{E}_{\theta|y}[\theta]\right)\right) \\ &= -2\left(\mathrm{E}_{\theta|y}\left[\log\left(\int f(y \mid \theta, \boldsymbol{\gamma}) P(\boldsymbol{\gamma} \mid \theta) d\boldsymbol{\gamma}\right)\right] - \log\left(\int f\left(y \mid \mathrm{E}_{\theta|y}[\theta], \boldsymbol{\gamma}\right) p\left(\boldsymbol{\gamma} \mid \mathrm{E}_{\theta|y}[\theta]\right) d\boldsymbol{\gamma}\right)\right) \end{aligned}$$

We use the subscript m to denote a marginalized p_D and henceforward we use the subscript j to denote what we will call a "joint" (or naive) p_D construction—that is, a p_D where the focus includes $\boldsymbol{\gamma}$ among the stochastic nodes of interest. Unfortunately, the value of p_D that many software packages calculate assumes γ to be a stochastic vector of interest, leading to the construction

$$p_{Dj} = -2\left(\mathrm{E}_{\theta,\boldsymbol{\gamma}|y}[\log f(y \mid \theta, \boldsymbol{\gamma})] - \log f\left(y \mid \mathrm{E}_{\theta,\boldsymbol{\gamma}|y}[\theta, \boldsymbol{\gamma}]\right)\right)$$

As the above example makes clear, there is a fundamental difference between p_{Dm} and p_{Dj}. Succinctly, the issues related to the application of p_D and DIC in instances like this are referred to as "the marginalization problem"—so named because the differences depend on whether or not γ is marginalized out before calculating p_D and DIC.

How big is this marginalization problem? Let us assume—uncharacteristically for this model, but it helps us to see an analytic example of the effect—that τ_g and τ_e are known. We also assume that ψ has an improper flat reference prior. With these assumptions, it is well known that

$$\begin{aligned}
\mathrm{E}[\psi \mid \mathbf{Y}] &= \frac{1}{N} J_N^T \mathbf{Y} && \equiv \overline{y}_{\cdot\cdot} \\
\mathrm{Var}[\psi \mid \mathbf{Y}] &= \frac{1}{N\tau_e} + \frac{1}{k\tau_g} && \equiv b
\end{aligned}$$

Further, let us define

$$\begin{aligned}
\overline{y}_{i\cdot} &= \frac{1}{n}\sum_{j=1}^{n} y_{ij} \\
\overline{Y} &= \begin{bmatrix} \overline{y}_{1\cdot} & \ldots & \overline{y}_{k\cdot} \end{bmatrix}^T \\
&= \left(\frac{1}{n} I_k \otimes J_n\right)^T \mathbf{Y}
\end{aligned}$$

Then we can write the distribution of the random effects, $\boldsymbol{\gamma}$, when $\mathbf{Y}$ and θ are known.

$$\gamma \mid \mathbf{Y}, \psi \sim \mathrm{N}_k\left(\frac{\tau_g}{\tau_g + n\tau_e}\psi J_k + \frac{n\tau_e}{\tau_g + n\tau_e}\overline{Y}, \frac{1}{\tau_g + n\tau_e} I_k\right).$$

We now give names to two quantities from the above distribution, to help us simplify our work below:

$$w \equiv \tau_g/(\tau_g + n\tau_e)$$

$$v \equiv 1/(\tau_g + n\tau_e)$$

This allows us to rewrite the above distribution of $\boldsymbol{\gamma}$ as

$$\gamma \mid \mathbf{Y}, \psi \sim \mathrm{N}_k\big(w\psi J_k + (1-w)\overline{Y}, vI_k\big).$$

Then plugging into the formula for p_{Dm}, we have

$$\begin{aligned}
&\mathrm{E}_{\theta|\mathbf{Y}}[\log f(y \mid \theta)] \\
&\quad = \mathrm{E}_{\psi|\mathbf{Y}}\left[\log\left((2\pi)^{-N/2}\left|\frac{1}{\tau_g} I_k \otimes J_n^n + \frac{1}{\tau_e} I_N\right|^{-1/2}\right) - \frac{1}{2}\left(\mathbf{Y} - \psi J_N\right)^T \left[\frac{1}{\tau_g} I_k \otimes J_n^n + \frac{1}{\tau_e} I_N\right]^{-1} \left(\mathbf{Y} - \psi J_N\right)\right] \\
&\quad = \log\left((2\pi)^{-N/2}\left|\frac{1}{\tau_g} I_k \otimes J_n^n + \frac{1}{\tau_e} I_N\right|^{-1/2}\right) - \mathrm{E}_{\psi|\mathbf{Y}}\left[\frac{1}{2}\left(\mathbf{Y} - \psi J_N\right)^T \left[\frac{1}{\tau_g} I_k \otimes J_n^n + \frac{1}{\tau_e} I_N\right]^{-1} \left(\mathbf{Y} - \psi J_N\right)\right]
\end{aligned}$$

for the first term, and

$$
\begin{aligned}
&\log f\left(y \mid \mathrm{E}_{\theta|\mathbf{Y}}[\theta]\right) \\
&\quad = \log\left((2\pi)^{-N/2}\left|\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right|^{-1/2}\right) \\
&\qquad - \frac{1}{2}\left(\mathbf{Y} - \mathrm{E}_{\psi|\mathbf{Y}}[\psi]\, J_N\right)^T \left[\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right]^{-1}\left(\mathbf{Y} - \mathrm{E}_{\psi|\mathbf{Y}}[\psi]\, J_N\right) \\
&\quad = \log\left((2\pi)^{-N/2}\left|\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right|^{-1/2}\right) - \frac{1}{2}\left(\mathbf{Y} - \overline{y}_{\cdot\cdot}J_N\right)^T \left[\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right]^{-1}\left(\mathbf{Y} - \overline{y}_{\cdot\cdot}J_N\right)
\end{aligned}
$$

for the second term. Since the constants of integration are the same and the logs of those constants cancel, combining these two terms leads to the following equation for p_{Dm}.

$$
\begin{aligned}
p_{Dm} &= \mathrm{E}_{\psi|\mathbf{Y}}\left[\left(\mathbf{Y} - \psi J_N\right)^T \left[\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right]^{-1}\left(\mathbf{Y} - \psi J_N\right)\right] \\
&\qquad - \left(\mathbf{Y} - \overline{y}_{\cdot\cdot}J_N\right)^T \left[\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right]^{-1}\left(\mathbf{Y} - \overline{y}_{\cdot\cdot}J_N\right) \\
&= \left(\mathbf{Y}^T \left[\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right]^{-1}\mathbf{Y} - \mathbf{Y}^T \left[\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right]^{-1}\mathbf{Y}\right) \\
&\qquad - \left(\mathrm{E}_{\psi|\mathbf{Y}}[\psi]\, J_N^T \left[\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right]^{-1}\mathbf{Y} - \overline{y}_{\cdot\cdot}J_N^T \left[\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right]^{-1}\mathbf{Y}\right) \\
&\qquad - \left(\mathbf{Y}^T \left[\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right]^{-1}\mathrm{E}_{\psi|\mathbf{Y}}[\psi]\, J_N - \mathbf{Y}^T \left[\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right]^{-1}\overline{y}_{\cdot\cdot}J_N\right) \\
&\qquad + \left(\mathrm{E}_{\psi|\mathbf{Y}}\left[\psi J_N^T \left[\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right]^{-1}\psi J_N\right] - \overline{y}_{\cdot\cdot}J_N^T \left[\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right]^{-1}\overline{y}_{\cdot\cdot}J_N\right) \\
&= \left(\mathrm{E}_{\psi|\mathbf{Y}}\left[\psi^2\right] - \overline{y}_{\cdot\cdot}^2\right) J_N^T \left[\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right]^{-1} J_N \\
&= \mathrm{Var}[\psi \mid \mathbf{Y}]\, J_N^T \left[\frac{1}{\tau_g}I_k \otimes J_n^n + \frac{1}{\tau_e}I_N\right]^{-1} J_N \\
&= b\tau_e \sum_{i=1}^{k} J_n^T \left[I_n - \frac{\tau_e}{\tau_g + n\tau_e}J_n^n\right] J_n \\
&= N\tau_e wb \\
&= 1
\end{aligned}
$$

So p_{Dm} in this setting is identically equal to 1, which is what we would like. The only free parameter in the model we have described is ψ—since τ_e and τ_g are both known; and although the random vector $\boldsymbol{\gamma}$ is a stochastic node in an MCMC sampler, it is not a parameter vector.

We now turn our attention to the calculation of p_{Dj}, the p_D construction with naive focus. Again, we start by specifying the elements of the p_{Dj} formula for this problem, given above.

The first term is given by

$$\begin{aligned}
&\mathrm{E}_{\theta,\boldsymbol{\gamma}|\mathbf{Y}}[\log f(y \mid \theta, \boldsymbol{\gamma})] \\
&\quad = \mathrm{E}_{\psi,\boldsymbol{\gamma}|\mathbf{Y}}\left[\log\left((2\pi)^{-N/2}\left|\frac{1}{\tau_e}I_N\right|^{-1/2}\right) - \frac{1}{2}\left(\mathbf{Y}-\boldsymbol{\gamma}\otimes J_n\right)^T\left[\frac{1}{\tau_e}I_N\right]^{-1}\left(\mathbf{Y}-\boldsymbol{\gamma}\otimes J_n\right)\right] \\
&\quad = \log\left((2\pi)^{-N/2}\left|\frac{1}{\tau_e}I_N\right|^{-1/2}\right) - \mathrm{E}_{\psi,\boldsymbol{\gamma}|\mathbf{Y}}\left[\frac{1}{2}\left(\mathbf{Y}-\boldsymbol{\gamma}\otimes J_n\right)^T\left[\frac{1}{\tau_e}I_N\right]^{-1}\left(\mathbf{Y}-\boldsymbol{\gamma}\otimes J_n\right)\right].
\end{aligned}$$

The second term is more complicated, necessitating our use of the Law of Total Expectation.

$$\begin{aligned}
&\log f\left(y \mid \mathrm{E}_{\theta,\boldsymbol{\gamma}|\mathbf{Y}}[\theta,\boldsymbol{\gamma}]\right) \\
&\quad = \log f\left(y \mid \mathrm{E}_{\psi|\mathbf{Y}}[\psi], \mathrm{E}_{\psi|\mathbf{Y}}\left[\mathrm{E}_{\boldsymbol{\gamma}|\mathbf{Y},\psi}[\boldsymbol{\gamma}]\right]\right) \\
&\quad = \log\left((2\pi)^{-N/2}\left|\frac{1}{\tau_e}I_N\right|^{-1/2}\right) \\
&\qquad\quad - \frac{1}{2}\left(\mathbf{Y}-\mathrm{E}_{\psi|\mathbf{Y}}\left[\mathrm{E}_{\boldsymbol{\gamma}|\mathbf{Y},\psi}[\boldsymbol{\gamma}]\right]\otimes J_n\right)^T\left[\frac{1}{\tau_e}I_N\right]^{-1}\left(\mathbf{Y}-\mathrm{E}_{\psi|\mathbf{Y}}\left[\mathrm{E}_{\boldsymbol{\gamma}|\mathbf{Y},\psi}[\boldsymbol{\gamma}]\right]\otimes J_n\right) \\
&\quad = \log\left((2\pi)^{-N/2}\left|\frac{1}{\tau_e}I_N\right|^{-1/2}\right) \\
&\qquad\quad - \frac{1}{2}\left(\mathbf{Y}-\left(w\overline{y}_{\cdot\cdot}J_N+(1-w)\overline{Y}\otimes J_n\right)\right)^T\left[\frac{1}{\tau_e}I_N\right]^{-1}\left(\mathbf{Y}-\left(w\overline{y}_{\cdot\cdot}J_N+(1-w)\overline{Y}\otimes J_n\right)\right).
\end{aligned}$$

Once again, we recognize that these terms have equal constants of integration, and that the logs of those constants cancel. Combining terms, we obtain the following equation for p_{Dj}.

$$\begin{aligned}
p_{Dj} &= \mathrm{E}_{\psi,\boldsymbol{\gamma}|\mathbf{Y}}\left[(\mathbf{Y}-\boldsymbol{\gamma}\otimes J_n)^T\left[\frac{1}{\tau_e}I_N\right]^{-1}(\mathbf{Y}-\boldsymbol{\gamma}\otimes J_n)\right] \\
&\qquad -\left(\mathbf{Y}-\left(w\overline{y}_{\cdot\cdot}J_N+(1-w)\overline{Y}\otimes J_n\right)\right)^T\left[\frac{1}{\tau_e}I_N\right]^{-1}\left(\mathbf{Y}-\left(w\overline{y}_{\cdot\cdot}J_N+(1-w)\overline{Y}\otimes J_n\right)\right) \\
&= \left(\mathbf{Y}^T\left[\frac{1}{\tau_e}I_N\right]^{-1}\mathbf{Y}-\mathbf{Y}^T\left[\frac{1}{\tau_e}I_N\right]^{-1}\mathbf{Y}\right) \\
&\qquad -\left(\left(\mathrm{E}_{\psi,\boldsymbol{\gamma}|\mathbf{Y}}[\boldsymbol{\gamma}]\otimes J_n\right)^T\left[\frac{1}{\tau_e}I_N\right]^{-1}\mathbf{Y}-\left(w\overline{y}_{\cdot\cdot}J_N+(1-w)\overline{Y}\otimes J_n\right)^T\left[\frac{1}{\tau_e}I_N\right]^{-1}\mathbf{Y}\right) \\
&\qquad -\left(\mathbf{Y}^T\left[\frac{1}{\tau_e}I_N\right]^{-1}\left(\mathrm{E}_{\psi,\boldsymbol{\gamma}|\mathbf{Y}}[\boldsymbol{\gamma}]\otimes J_n\right)-\mathbf{Y}^T\left[\frac{1}{\tau_e}I_N\right]^{-1}\left(w\overline{y}_{\cdot\cdot}J_N+(1-w)\overline{Y}\otimes J_n\right)\right) \\
&\qquad +\mathrm{E}_{\psi,\boldsymbol{\gamma}|\mathbf{Y}}\left[(\boldsymbol{\gamma}\otimes J_n)^T\left[\frac{1}{\tau_e}I_N\right]^{-1}(\boldsymbol{\gamma}\otimes J_n)\right] \\
&\qquad -\left(w\overline{y}_{\cdot\cdot}J_N+(1-w)\overline{Y}\otimes J_n\right)^T\left[\frac{1}{\tau_e}I_N\right]^{-1}\left(w\overline{y}_{\cdot\cdot}J_N+(1-w)\overline{Y}\otimes J_n\right)
\end{aligned}$$

Note that each of the first three lines of the foregoing equality are equal to 0. Then

$$\begin{aligned}
p_{Dj} &= \tau_e\left(\mathrm{E}_{\psi,\boldsymbol{\gamma}|\mathbf{Y}}\left[(\boldsymbol{\gamma}\otimes J_n)^T\boldsymbol{\gamma}\otimes J_n\right]-\left(w\overline{y}_{\cdot\cdot}J_N+(1-w)\overline{Y}\otimes J_n\right)^T\left(w\overline{y}_{\cdot\cdot}J_N+(1-w)\overline{Y}\otimes J_n\right)\right) \\
&= n\tau_e\left(\mathrm{E}_{\psi,\boldsymbol{\gamma}|\mathbf{Y}}\left[\boldsymbol{\gamma}^T\boldsymbol{\gamma}\right]-\left(w\overline{y}_{\cdot\cdot}J_k+(1-w)\overline{Y}\right)^T\left(w\overline{y}_{\cdot\cdot}J_k+(1-w)\overline{Y}\right)\right) \\
&= n\tau_e\left(\mathrm{E}_{\psi|\mathbf{Y}}\left[\sum_{i=1}^{k}\mathrm{E}_{\boldsymbol{\gamma}_i|\mathbf{Y},\psi}\left[\boldsymbol{\gamma}_i^2\right]\right]-\sum_{i=1}^{k}(w\overline{y}_{\cdot\cdot}+(1-w)\overline{y}_{i\cdot})^2\right) \\
&= n\tau_e\sum_{i=1}^{k}\left(\mathrm{E}_{\psi|\mathbf{Y}}\left[\mathrm{Var}_{\gamma_i|\mathbf{Y},\psi}[\gamma_i]+\mathrm{E}_{\gamma_i|\mathbf{Y},\psi}[\gamma_i]^2\right]-(w\overline{y}_{\cdot\cdot}+(1-w)\overline{y}_{i\cdot})^2\right)
\end{aligned}$$

Recalling that $\gamma \mid \mathbf{Y},\psi \sim \mathrm{N}_k\left(w\psi J_k+(1-w)\overline{Y},vI_k\right)$, we can now finish simplifying the equation for p_{Dj} in this setting.

$$
\begin{aligned}
p_{Dj} &= n\tau_e \sum_{i=1}^{k} \left(\mathrm{E}_{\psi|\mathbf{Y}}\left[v + (w\psi + (1-w)\overline{y}_{i\cdot})^2 \right] - (w\overline{y}_{\cdot\cdot} + (1-w)\overline{y}_{i\cdot})^2 \right) \\
&= N\tau_e v + n\tau_e \sum_{i=1}^{k} \left(\mathrm{E}_{\psi|\mathbf{Y}}\left[w^2\psi^2 + 2w(1-w)\overline{y}_{i\cdot}\psi + (1-w)^2\overline{y}_{i\cdot}^2 \right] - (w\overline{y}_{\cdot\cdot} + (1-w)\overline{y}_{i\cdot})^2 \right) \\
&= N\tau_e v + n\tau_e \sum_{i=1}^{k} \left(w^2 \left(\mathrm{E}_{\psi|\mathbf{Y}}\left[\psi^2\right] - \overline{y}_{\cdot\cdot}^2 \right) + 2w(1-w)\left(\overline{y}_{i\cdot}\,\mathrm{E}_{\psi|\mathbf{Y}}[\psi] - \overline{y_{i\cdot}}^2 \right) + (1-w)^2\left(\overline{y_{i\cdot}}^2 - \overline{y_{i\cdot}}^2 \right) \right) \\
&= N\tau_e v + n\tau_e \sum_{i=1}^{k} \left(w^2 \,\mathrm{Var}_{\psi|\mathbf{Y}}[\psi] \right) \\
&= N\tau_e(v + w^2 b) \\
&= 1 + (k-1)\frac{n\tau_e}{\tau_g + n\tau_e}
\end{aligned}
$$

As we saw, p_{Dm} is identically equal to 1 in this setting. Similarly, when τ_g is much larger than $n\tau_e$, p_{Dj} also approaches 1. When τ_g is much smaller than $n\tau_e$, however, p_{Dj} approaches k. This is reasonable given the choice of focus and the situation described. The quantity p_{Dm} identifies a single free parameter, ψ. When $\tau_g \gg n\tau_e$, there is very little variability in the random effects terms relative to the variability within individuals, and the data behave like they come from a common population and p_{Dj} is near 1. When $\tau_g \ll n\tau_e$ on the other hand, the data behave like k separate populations, each having their own effect.

This, then, is the marginalization problem. Although the marginal construction gives $p_D = 1$, as we would expect, the naive construction gives $1 \leq p_D \leq k$. In the next section, we discuss why we believe this inconsistency necessitates use of the marginal construction.

3.3 The Need for Marginalization

We will now endeavor to describe three distinct constructions for p_D and DIC in the mixed model setting. These are the joint and marginal constructions, as discussed above, and the

BUGS numerical approximations. After discussing these three constructions, we proceed to give our argument as to why we believe the marginal construction should be preferred. Finally, we discuss some counterarguments against a preference for marginalization in model selection.

3.3.1 Three DIC Constructions for Mixed Models

In the following three chapters, we make extended reference to three different numerical approximations to DIC: the joint DIC, the BUGS DIC, and the marginal DIC. As explained above, in mixed models the value of the DIC depends on the choice of focal objects. We now explain the difference between these three approximations, and why we consider the marginal DIC to be philosophically preferable for doing model selection in the mixed model setting.

3.3.1.1 The joint DIC

The first construction for DIC is what one might consider the naive construction. This construction assumes that all stochastic objects are of focal interest. It centers on what we call a joint likelihood for both θ and γ,

$$L(\theta, \gamma \mid y) \propto f(y \mid \gamma, \theta).$$

Under the joint construction, we have the following definitions:

$$\begin{aligned} p_{Dj} &= \mathrm{E}_{\theta,\gamma|y}[-2\log L(\theta,\gamma \mid y)] + 2\log L(\hat{\theta},\hat{\gamma} \mid y) \\ &= \overline{D(\theta,\gamma)} - D(\hat{\theta},\hat{\gamma}), \\ DIC_j &= 2\,\mathrm{E}_{\theta,\gamma|y}[-2\log L(\theta,\gamma \mid y)] + 2\log L(\hat{\theta},\hat{\gamma} \mid y) \\ &= 2\overline{D(\theta,\gamma)} - D(\hat{\theta},\hat{\gamma}). \end{aligned}$$

In other words, the joint DIC treats γ as if it were a model parameter alongside θ, and uses the posterior mean of both θ and γ to obtain the fitted deviance. Here, $\hat{\theta} = \mathrm{E}_{\theta,\gamma|y}[\theta]$ and $\hat{\gamma} = \mathrm{E}_{\theta,\gamma|y}[\gamma]$.

The joint construction above is similar to the construction for DIC_7 in Celeux et al. (2006), with two notable differences. First, they choose $\hat{\theta}$ and $\hat{\gamma}$ to be the joint maximum *a posteriori* (MAP) estimates conditional on y. We choose $\hat{\theta}$ and $\hat{\gamma}$ to be the posterior means. Their choice to use joint MAP estimates is based on the poor behavior of estimators of γ in latent random variable problems and their concern with how the DIC behaves in mixture models, where the posterior mean may live in an area of relatively low posterior density. Our use of the posterior mean stems from SBCV's recommendation to use posterior means and our belief that this construction is more likely to be used than DIC_7 by others who might encounter a mixed modeling scenario.

Second, while we call this a joint construction, Celeux et al. (2006) call it a conditional construction. This is a fundamental difference in how we regard these constructions. We call this a joint construction because θ and γ appear jointly in a likelihood statement, and are considered jointly by the DIC formulae. They call this a conditional construction because the density (as opposed to likelihood) of interest is $f(y \mid \gamma, \theta)$, where y is conditioned on γ.

3.3.1.2 The BUGS DIC

Our information on how OpenBUGS constructs the DIC is drawn from *The BUGS Book* (Lunn et al. 2013) and the *OpenBUGS User Manual* (D. Spiegelhalter et al. 2014). The manual describes how OpenBUGS obtains $\overline{D(\theta)}$ and $D(\hat{\theta})$ as follows:

> $\overline{D(\theta)}$: this is the posterior mean of the deviance, which is exactly the same as if the node 'deviance' had been monitored. This deviance is defined as $-2\log P(y|\theta)$, where y comprises all stochastic nodes given values (i.e. data), and θ comprises the stochastic parents of y – 'stochastic parents' are the stochastic nodes upon which the distribution of y depends, when collapsing over all logical relationships.
>
> $D(\hat{\theta})$: this is a point estimate of the deviance ($-2\log P(y|\theta)$) obtained by substituting in the posterior means $\frac{1}{B}\sum_{s=1}^{B}\theta^{(s)}$ of θ: thus $D(\hat{\theta}) = -2\log p\left(y \mid \frac{1}{B}\sum_{s=1}^{B}\theta^{(s)}\right)$

As this construction pertains to hierarchical models, *The BUGS Book* describes numerical approximations to DIC in the BUGS family of programs as follows:

> WinBUGS (and OpenBUGS) separately reports the contribution to $\overline{D(\theta)}$, p_D, and DIC for each differently named (scalar, vector, or array) node, together with a total. This enables the individual contributions from different portions of data to be assessed. In some circumstances some of these contributions may need to be ignored and removed from the total.

This is not, unfortunately, enough information to classify OpenBUGS's construction of the DIC for hierarchical models into the framework provided by Celeux et al. (2006). We do not have sufficient information on which nodes constitute stochastic parents of our data in the mixed model. We can state, however, that for all the models considered in this dissertation, OpenBUGS reports DIC contributions from only our data y and cannot be partialed out

as described above. We believe, based on the information presented in the manual and *The BUGS Book*, that OpenBUGS's construction of the DIC should match our DIC_j, but simulation results presented in Chapters 4 and 5 confirm that there are differences in the numerical approximation algorithms.

We nonetheless present results for DIC_b and related quantities because we consider it important to compare the numerical approximations under methods we develop to what is given by commercially available software, since numerical approximations to DIC given by commercially available software are what practitioners are most likely to use.

3.3.1.3 The marginal DIC

Our final construction focuses only on the stochastic node θ, treating γ as a latent random vector to be marginalized out. The marginal DIC can be expressed using three different likelihood functions, which we briefly clarify before giving the formulae for this construction.

$$
\begin{aligned}
L(\theta \mid y, \gamma) &\propto f(y, \gamma \mid \theta) \\
L(\theta, \gamma \mid y) &\propto f(y \mid \gamma, \theta) \\
L(\theta \mid y) &\propto f(y \mid \theta) \\
&= \int f(y, \gamma \mid \theta) d\gamma \\
&= \int f(y \mid \gamma, \theta) P(\gamma \mid \theta) d\gamma \\
&= \mathrm{E}_{\gamma \mid \theta}[L(\theta, \gamma \mid y)]
\end{aligned}
$$

Then we define the marginal construction with a few equivalent expressions:

$$\begin{aligned}
p_{Dm} &= \mathrm{E}_{\theta|y}[-2\log L(\theta \mid y)] + 2\log L(\hat{\theta} \mid y) \\
&= \mathrm{E}_{\theta|y}\left[-2\log \int L(\theta \mid y, \gamma) d\gamma\right] + 2\log \int L(\hat{\theta} \mid y, \gamma) d\gamma \\
&= \mathrm{E}_{\theta|y}\left[-2\log \mathrm{E}_{\gamma|\theta}[L(\theta, \gamma \mid y)]\right] + 2\log \mathrm{E}_{\gamma|\theta}\left[L(\hat{\theta}, \gamma \mid y)\right] \\
&= \overline{D(\theta)} - D(\hat{\theta}) \\
DIC_m &= 2\,\mathrm{E}_{\theta|y}[-2\log L(\theta \mid y)] + 2\log L(\hat{\theta} \mid y) \\
&= 2\overline{D(\theta)} - D(\hat{\theta}).
\end{aligned}$$

where

$$\begin{aligned}
\hat{\theta} &= \mathrm{E}_{\theta|y}[\theta] \\
&= \int \theta P(\theta \mid y) d\theta \\
&= \int \theta \int P(\theta, \gamma \mid y) d\gamma d\theta \\
&= \int \int \theta P(\theta, \gamma \mid y) d\gamma d\theta \\
&= \mathrm{E}_{\theta,\gamma|y}[\theta]
\end{aligned}$$

Note that, as shown above, the quantities in the marginal construction can be written both as integrals of likelihoods and as expectations over the distribution $P(\gamma \mid \theta)$. These are subtly different interpretations, and both will prove useful to us in our discussion of methods for approximating DIC_m in Section 4.2.2.

This construction is given by Celeux et al. (2006) as DIC_1, who refer to it as an "observed DIC" to match their terminology for $L(\theta \mid y)$, which they call an observed likelihood.

To simplify notation where necessary, we use the j subscript (e.g. p_{Dj}, DIC_j, $\overline{D}_j$, $D(\hat{\theta})_j$) to refer to the DIC calculations focusing on the joint distribution of θ and γ. Similarly, we use the b subscript for the BUGS calculations and the m subscript for the marginal calculations.

3.3.2 Why Do We Prefer the Marginal DIC?

Having established that p_D can depend the set of focal stochastic objects with our example in Section 3.2.2, we now consider whether this dependence is worth our concern. We believe so, and in the subsections below we make our case for using the marginal DIC. We give three reasons based on: (1) the conceptual difference between adopting the marginal or the joint foci, (2) the rise of automated model selection procedures, (3) and the interpretation of p_D in hierarchical models.

3.3.2.1 Conceptual differences between marginalized and joint foci

We begin with a discussion of what it means for the DIC to focus on θ and γ, rather than θ alone. An individual-level random effect can, in general, be thought of as a catch-all correction factor encapsulating all of the remaining differences among individuals that are germane, after conditioning on every covariate already measured and included in the model. For example, if we consider the human fertility study referenced in the preceding chapter, a model might suggest that the probability of ovulation during a particular cycle is a function of certain covariates: e.g. ethnicity, age, weight, average daily caffeine intake, etc. But information on these covariates alone may not be sufficient to describe the differences observed among individuals. There may also be additional random slopes—individual differences in the relationships between time-varying covariates and the response of interest.

Random effects are usually specified by a particular parametric distribution, and the random effect for each individual is assumed to be independently drawn from this distribution. Researchers may be interested in the parameters of this random effects distribution—measures that reflect how much inter-individual variability remains in the response data that hasn't been captured by the covariates[3]. These so-called variance components are included in θ. The (random) γ's are the latent random effects, which under such a distribution reflect how far an individual's response differs from the overall population mean, adjusted for observed covariates.

The choice of whether one is interested in making inferences for particular γ's is essentially a choice about whether one is concerned with the population of individuals who haven't been sampled, or concerned only with the individuals in the sample. Both choices can be reasonable—but when one is concerned only with understanding the individuals in the sample, this is more accurately reflected by considering a fixed effect for those individuals. The choice to consider random, rather than fixed, effects is essentially a choice to prioritize generalizability. Otherwise why would one be concerned with the distributional properties of the random effects?

We consider that the marginalized approach to be philosophically preferable. The real distinction between fixed and random effects is whether one wants to make specific inferences about the observed clusters in particular, or whether one wants to extrapolate to the general population from which those clusters were sampled. If one wants to make inferences about the observed clusters, then one should fit a fixed effects model. The usual DIC, in that case, requires no marginalization. If one doesn't care about observed clusters, then the only parameter of interest should be the covariance matrix for the random effects, $\mathbf{\Sigma}$. This leads

[3]Note that the example in Section 3.2.2 was developed under the assumption that these values were known. This was done in order to provide insight about the behavior of p_D as a function of the precisions for the random effects and error distributions. SBCV consider a similar example with ψ constant and τ_e unknown, obtaining analogous results.

to the need to marginalize over the γ and simply focus on the parameters of interest, which includes $\mathbf{\Sigma}$ as well as any other model parameters.

Suggestions have been made to us that a scientist might be concerned with both generalizability to a wider population, *and* with details of individual-level effects for the sampled clusters. Some may have such interests—but considering the role of DIC as a model selection criterion, for it to be meaningful, a prioritization must be made. We have observed that p_D can depend on the choice of focal stochastic objects, and that the θ and (θ, γ) foci can yield different results. A fixed-effects structure replacing random effects gives results that differ from either of these, since the parameters of the random effects distribution are not of focal interest when those effects are considered as fixed. What we are left with, then, is the choice between three possible DIC calculations. A scientist whose interests are only out-of-sample generalizability should use DIC_m. A scientist whose interests are only on the k sampled clusters should use the fixed effects model and its associated DIC. We believe that the third option, the DIC_j construct, is never preferable to these. It depends on the ratio of random effects and error variances, and its meaning in the mixed model setting remains unclear. Certainly, it does not appear possible to argue that DIC_j represents a principled reweighting of DIC_m and the fixed effects DIC that will always reflect the inferential priorities of the user.

3.3.2.2 Automated model selection requires carefully chosen tools

The issue of focal choice is further complicated by the increasing reliance on automated model selection procedures. As a thought example, consider how Google places advertisements on websites. The following information is condensed and summarized from Google's AdSense Help Center (Help 2017).

When a website using Google AdSense has adspace for sale, computers at Google classify the website according to "factors [such] as keyword analysis, word frequency, font size, and the overall link structure of the web." Then the Google systems search a database of advertisers and select those whose ads are deemed relevant to the content or users of the website. Google creates an automated auction where advertisers can bid on the available adspace in units of cost-per-click (CPC), which is how much the advertiser is willing to pay the website owner for each click their advertisement receives. Google combines CPC bids with a quality score—a measure of how likely an ad is to be clicked based on its past performance and how well its content matches the website—to decide which advertiser wins an auction. Further, Google estimates how likely it is that an ad click will lead to a business transaction for its advertisers, and dynamically reduces some advertiser bids. The rationale behind this practice is that it protects advertisers from overspending on advertisements that are unlikely to result in business transactions, and allows advertisers to bid more freely in the auctions.

A statistician will recognize many areas in this process where covariate modeling and model selection procedures are relevant. Which website factors will best predict click-through rate (CTR) for a particular ad or class of ads? How much should an advertiser bid in a certain situation if that advertiser wants its ads to be seen? Which ad and advertiser characteristics best predict that ad clicks will result in business transactions? Because of the speed and frequency necessary for these decision-making problems, however, direct supervision is difficult if not impossible. New advertisements, and new websites, enter the marketplace too quickly for individual analysts to study or classify them. Simplicity of classification will tend to result in less content-specific ad placement, reducing revenue for the website owner, the advertisers, and Google itself. Incentives are high, in this situation, to create model selection algorithms that do not need supervision. This is an example of the discipline of machine learning.

Modern statisticians must anticipate encountering situations where it is necessary to choose a principled model selection procedure that behaves in a desired fashion even without close monitoring. With the rise of "big data," it is now more important than ever that statisticians and scientists have a clear understanding of their model selection tools—especially how those tools may give different results than other model selection tools, and which tool selects models that are preferred for a given application. Even in relatively simple settings, the various model selection criteria discussed above can lead to considerably different model choices. R. R. Christensen (2017) has shown that when comparing two nested linear models, selection by Adjusted R^2 is equivalent to selecting the larger model when the F statistic is greater than 1, selection by C_p corresponds to $F > 2$, selection by AIC is asymptotically equivalent to $F > 2$, and selection by BIC corresponds asymptotically to $F > \log n$. For more complicated settings like LMMs and GLMMs, good understanding of available criteria is even more important. We consider this another reason why the *marginal* DIC calculations should be preferred to other DIC calculation methods. Marginal DIC calculations are more easily understood, because the theory surrounding them is relatively straightforward compared to the broader theory surrounding DIC calculations for hierarchical models (c.f. Celeux et al. 2006). The asymptotic equivalence we showed between DIC and AIC in Section 3.1.7 fails when random effects are included in the model.

3.3.2.3 Interpreting p_D in hierarchical models

The problem of using DIC_j for model selection in hierarchical models has received considerable attention, especially as it relates to p_D, and has already been discussed by us in Section 3.2.2. Brooks (2002) explains that "[s]adly, in many cases the calculation of p_D will be impossible for the focus of primary interest since the deviance will not be available in closed [form]," including in random effects and state-space models. To elucidate the behavior of p_D, Sahu (2002) provides a simpler version of our own example in Section 3.2.2 to discuss

the fact that $p_D \to \infty$ as $k \to \infty$ under the DIC_j construct. Although not related directly to the DIC, Su and Johnson (2006) provide contributions explaining the asymptotic behavior of random effects models with respect to the roles of large n and large k.

Celeux et al. (2006) provide a comprehensive review of a number of DIC constructions and associated issues. As we mentioned above, our DIC_m corresponds directly with their DIC_1, and our DIC_j roughly with their DIC_7. Celeux et al. are concerned with cases such as mixture models where $\mathrm{E}[\theta \mid y, \gamma]$ may result in poor performance of DIC leading to a negative p_D. They show that in certain problems, using the maximum *a posteriori* (MAP) estimates for θ and γ can result in better behavior than the posterior mean. Celeux et al. also concur with our assessment that constructions of the DIC_j form, those that treat the latent random variable γ like a parameter vector, are unsatisfactory. They state that "this approach has obvious asymptotic and coherency difficulties, as discussed in previous literature" and "in the random effect model... computing the p_D's and therefore the DIC's does not really make sense."

In this section, we have argued that when random effects are needed, DIC_m is the sensible construct to consider because it correctly treats the random effects as "nuisance". We have argued that understanding the behavior of model selection criteria is especially important in situations where model selection must be automated, and that we should avoid criteria whose behavior is difficult to understand. And we have discussed the concerns other researchers have expressed with the behavior of p_D in hierarchical models. Neither DIC_j nor—based on their reported numerical results—any of the other constructions considered by Celeux et al. (2006) provide estimates of the number of parameters in an hierarchical model that match the number we would expect from a marginalized model.

3.3.3 Arguments Against Marginalization

We have made our case for why we believe the marginal definition for DIC is to be preferred. We recognize, however, that our position is not universally held. Below, we discuss two critiques of the marginal preference that we have encountered.

3.3.3.1 Criterion instability with variance components

We have been made aware of inconsistencies in criterion behavior when selecting among models with different variance component structures. Specifically, Dr. Daniel Gillen of the University of California, Irvine, has mentioned that information criteria can exhibit a "skipping" behavior when variance components are added to or removed from a model. We believe this may be analogous to an effect we have previously observed in our own work with Dr. Gillen, which involved in part the simulation of a linear mixed effects models with a LASSO penalty. The simulation behavior of LASSO models is often evaluated in terms of out-of-sample prediction error as the LASSO penalty, λ, varies. In our work with Dr. Gillen, we observed that in mixed effects models, the prediction error for five-fold cross-validation as a function of λ was not a continuous function for linear mixed models; it generally does appear as a continuous function for fixed effect models. Figure 3.1, taken from this earlier simulation work, displays the skipping behavior we describe to help the reader envision the phenomenon.

This skipping behavior occurs when the LASSO adds or removes a covariate. When a covariate is added or removed, assuming this covariate relates to the response variable, the random effect appears to lose or gain (respectively) variability to account for the change in the fixed effects model. This assumes, of course, that the random effects are also related to the covariates.

Figure 3.1: "Skipping" behavior in a simulation study of LASSO use for linear mixed models. As the LASSO penalty (λ) increases, five-fold cross-validation prediction error makes distinct jumps at certain lambda values.

Dr. Gillen reports that similar phenomena can be observed when information criteria are used for model selection purposes and variance components are allowed to be added and removed as in the marginal selection setting we describe. Our examples in this dissertation all presuppose a cluster-level random effect and do not appear to be subject to these issues; but studies that involve multi-level clustering (c.f. the cow abortion data of Thurmond et al. (2005) that we will describe in Section 6.2.3) may require us to choose which clusters are and are not modeled with random effects distributions. Above, we advocated for the use of selection criteria whose behavior is well-understood and consistent, especially in automated selection settings. We continue to advocate this policy here as well, and believe that further investigation of the behavior of marginal selection criteria like DIC_m is warranted when selecting among variance components.

3.3.3.2 Inferential priorities

Some statisticians suggest that model selection for mixed models should take account of both the random effects for the observed clusters and the variance components for those random effects. They suggest this because scientific interest can reside in both areas simultaneously: how the model functions for new observations in the sampled clusters, as well as how it functions for new observations on new clusters. In the Bayesian setting, fixed effects and random effects have very similar specifications within a probability model; the primary difference between them is how they are handled in inference, once a posterior sample has been obtained. We agree that both the conditional effects on a response when cluster is known, and the marginal effects on a response are legitimate areas of scientific interest, but as discussed above we find it difficult to carefully define how model selection should proceed when both conditional and marginal inferences are desired.

Nonetheless, the argument has been made to us that, following from the example in Section 3.2.2 one should reasonably want to penalize a model as if it has k fixed effects if data are

sufficiently different from cluster to cluster; or that one should penalize a model as if it had only a grand mean if data are sufficiently homogenous across clusters. "The behavior of p_{Dj} is not a bug, it's a feature," one might say. This is a view we have encountered with some frequency, though we remain troubled by the fact that this argument presumes that the appropriate penalization of a model is in some sense dependent on the number of clusters one happens to select, even when inference for those clusters is not itself desired.

Other statisticians suggest that DIC may not be the preferred tool for situations such as the ones we describe. Gelman et al. (2013) place DIC in a hierarchy with AIC and BIC where they suggest that DIC should be preferred when inference is desired for individuals within the sampled clusters, AIC should be preferred when inference is desired for unsampled clusters of a similar character, and BIC should be preferred when inference is desired marginally on the population. This suggests that they consider DIC less useful for model selection relative to marginal population-level models, though we believe the marginalization techniques we develop in this dissertation broaden the scope of situations in which DIC may be usefully applied.

3.4 Marginalization in the Linear Mixed Model

Our arguments in the preceding section lead to the question of why marginalization is not performed more often when selecting a model. One answer is that marginalization is difficult, especially in the GLMM setting where closed-form marginal equations do not exist. Marginalization is both possible and practical in the LMM setting, however, and so we begin by explaining two marginalization approaches for the DIC. The methods explained here, particularly our approach to postprocessing marginalization, point the way toward the methods we develop in the next two chapters for marginalization of GLMMs.

3.4.1 Methods for Marginalization

In the linear mixed model, when a normal distribution is used for the random effects, there are two methods for getting the marginal likelihood $L(\theta \mid y)$ and thus the marginal DIC calculations. The first, which we call the preprocessing approach, involves expressing the model directly in its marginalized form. MCMC sampling directly using the marginal model will, obviously, yield the desired marginal DIC calculations. The second approach, which we refer to as postprocessing marginalization, uses the complete-the-square formula (Proposition 3.1, below) after re-expressing $f(y, \gamma \mid \theta)$ as $f(\gamma \mid y, \theta) f(y \mid \theta)$ when both $\gamma \mid \theta$ and $y \mid \gamma, \theta$ have multivariate normal distributions.

3.4.1.1 Preprocessing marginalization

In this section we discuss MCMC sampling that is based on using the marginal likelihood for θ, $L(\theta \mid y) \propto \int f(y \mid \gamma, \theta) P(\gamma \mid \theta) d\gamma$. When the marginal likelihood is available in closed form, as is the case for the LMM discussed above, it is relatively straightforward to implement MCMC sampling to obtain iterates $\{\theta^{(1)}, ..., \theta^{(B)}\}$ from the posterior, $P(\theta \mid y)$, with the help of packaged software like OpenBUGS, JAGS, STAN etc. This will involve monitoring $D(\theta) = -2\log(L(\theta \mid y))$ in one of these packages, to obtain $\overline{D(\theta)}$. Then, we use $\hat{\theta} = \frac{1}{B}\sum_{s=1}^{B} \theta^{(s)}$ to numerically approximate p_{Dm} and DIC_m with

$$p_{Dm} \simeq -\frac{2}{B}\sum_{s=1}^{B} \log f\left(Y \mid \theta^{(s)}\right) + 2\log f\left(Y \mid \hat{\theta}\right)$$
$$DIC_m \simeq -\frac{4}{B}\sum_{s=1}^{B} \log f\left(Y \mid \theta^{(s)}\right) + 2\log f\left(Y \mid \hat{\theta}\right)$$

We now proceed to analytically obtain the marginal likelihood.

Refer to Equation (3.9), the matrix specification for the linear mixed model. We can rewrite this model as $\mathbf{Y} - \mathbf{X}\beta = \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e}$. Recall that $\boldsymbol{\gamma} \sim \mathrm{N}_{kq}(\boldsymbol{\psi}, \boldsymbol{\Sigma})$ and $\mathbf{e} \sim \mathrm{N}_{kn}(0_{kn}, \sigma^2 I_{kn})$. Then we

can write the marginal for the data as

$$\mathbf{Y} \sim \mathrm{N}_{kn}\Big(\mathbf{X}\beta + \mathbf{Z}\boldsymbol{\psi}, \mathbf{Z}^T\boldsymbol{\Sigma}\mathbf{Z} + \sigma^2 I_{kn}\Big). \tag{3.10}$$

Since the matrix Σ_i is often relatively uncomplicated—in the case of a random intercepts model, it is the scalar variance of the random intercepts—it is thus easy to write the linear mixed model in terms of its induced marginal mean vector and covariance matrix, ignoring $\boldsymbol{\gamma}$ entirely.

If this approach is used, we should be cognizant of how MCMC sampling efficiency is affected. MCMC sampling for Bayesian models can be improved by including intermediate stochastic nodes like $\boldsymbol{\gamma}$, so avoiding them as we do in the preprocessing approach may lead to sampling behaviors we dislike. Preprocessing ensures that the numerical DIC approximations obtained from software like OpenBUGS are approximations to the desired DIC_m construct, but we must weigh this against the potential loss of sampling efficiency under this approach.

3.4.1.2 Postprocessing marginalization

Here we begin with the full description of the model that involves γ. We sample from this model involving γ, but then obtain an analytical form for the marginal into which we can plug our posterior iterates to obtain our own numerical approximation to DIC_m.

We start by writing the joint density for the data and γ conditional on θ, $f(y, \gamma \mid \theta)$, and then through a series of algebraic manipulations, we obtain an equivalent expression, namely $f(y, \gamma \mid \theta) = f(y \mid \theta) f(\gamma \mid y, \theta)$, where the conditional distribution in γ is normal with parameters depending on θ. Thus upon integrating over γ, we obtain an analytical expression for $f(y \mid \theta)$. Thus given a MC sample from the posterior for θ, which is easily obtained using BUGS or some other package, we are able to numerically approximate the marginal model based DIC.

Below, we develop a new expression for the marginal density $f(y \mid \theta)$. This is not necessary for the LMM setting—it should be clear that the marginal form in Equation (3.10) will serve this purpose, and in fact must be equivalent to the expression we develop below. The work we present here is crucial to subsequent work in our development of a marginalization approach for GLMMs during the next two chapters. In the GLMM setting, no closed-form marginalization exists and we are unable to use a preprocessing approach. We thus consider it preferable to introduce this work here where there are no complications.

Our alternate expression for the marginal density takes advantage of the fact that both $f(y \mid \gamma, \theta)$ and $P(\gamma \mid \theta)$ are normal densities. Mathematically, we can use the complete-the-square formula to combine the two and isolate the γ terms.

Following from the notation in Section 3.2.1, we assume $\mathbf{X}$ and $\mathbf{Z}$ are full rank and write the following:

$$
\begin{aligned}
f(y_{ij} \mid \gamma_i, \theta) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{y_{ij} - (x_{ij}\beta + z_{ij}\gamma_i)}{\sigma}\right)^2\right) \\
f(Y_i \mid \gamma_i, \theta) &= (2\pi\sigma^2)^{-k/2} \exp\left(-\frac{1}{2}\left[Y_i - (X_i\beta + Z_i\gamma_i)\right]^T \left(\sigma^2 I_n\right)^{-1} \left[Y_i - (X_i\beta + Z_i\gamma_i)\right]\right) \\
f(\gamma_i \mid \theta) &= (2\pi)^{-q/2} |\Sigma_i|^{-1/2} \exp\left(-\frac{1}{2}\left[\gamma_i - \psi_i\right]^T \Sigma_i^{-1} \left[\gamma_i - \psi_i\right]\right)
\end{aligned}
$$

This gives the joint density

$$
\begin{aligned}
& f(Y_i, \gamma_i \mid \theta) \\
&\quad = (2\pi)^{-(k+q)/2} \sigma^{-k} |\Sigma_i|^{-1/2} \\
&\qquad \times \exp\left(-\frac{1}{2}\left[Y_i - (X_i\beta + Z_i\gamma_i)\right]^T \left(\sigma^2 I_n\right)^{-1} \left[Y_i - (X_i\beta + Z_i\gamma_i)\right]\right) \\
&\qquad \times \exp\left(-\frac{1}{2}\left[\gamma_i - \psi_i\right]^T \Sigma_i^{-1} \left[\gamma_i - \psi_i\right]\right).
\end{aligned}
$$

To simplify notation, define $\tilde{Y}_i = Y_i - X_i\beta$. This allows us to write

$$\begin{aligned}
&\left[\tilde{Y}_i - Z_i\gamma_i\right]^T \left(\sigma^2 I_n\right)^{-1} \left[\tilde{Y}_i - Z_i\gamma_i\right] \\
&= \left[\tilde{Y}_i - Z_i\hat{\gamma}_i + Z_i\hat{\gamma}_i - Z_i\gamma_i\right]^T \left(\frac{1}{\sigma^2} I_n\right) \left[\tilde{Y}_i - Z_i\hat{\gamma}_i + Z_i\hat{\gamma}_i - Z_i\gamma_i\right] \\
&= \left[\tilde{Y}_i - Z_i\hat{\gamma}_i\right]^T \left(\frac{1}{\sigma^2} I_n\right) \left[\tilde{Y}_i - Z_i\hat{\gamma}_i\right] + \left[Z_i\hat{\gamma}_i - Z_i\gamma_i\right]^T \left(\frac{1}{\sigma^2} I_n\right) \left[Z_i\hat{\gamma}_i - Z_i\gamma_i\right] \\
&\qquad + \left[\tilde{Y}_i - Z_i\hat{\gamma}_i\right]^T \left(\frac{1}{\sigma^2} I_n\right) \left[Z_i\hat{\gamma}_i - Z_i\gamma_i\right] + \left[Z_i\hat{\gamma}_i - Z_i\gamma_i\right]^T \left(\frac{1}{\sigma^2} I_n\right) \left[\tilde{Y}_i - Z_i\hat{\gamma}_i\right] \\
&= \left[\tilde{Y}_i - Z_i\hat{\gamma}_i\right]^T \left(\frac{1}{\sigma^2} I_n\right) \left[\tilde{Y}_i - Z_i\hat{\gamma}_i\right] + \left[\hat{\gamma}_i - \gamma_i\right]^T \left(\frac{1}{\sigma^2} Z_i^T Z_i\right) \left[\hat{\gamma}_i - \gamma_i\right].
\end{aligned}$$

Then apply the "complete-the-square" formula below, which is proven in the appendix, to combine the quadratic terms for γ_i in the exponent. This results in a Normal kernel for γ_i and a second term that is free of γ_i, making it possible for us to easily marginalize.

Proposition 3.1. *For conformable vectors X, μ_1, and μ_2; and for conformable symmetric matrices A_1 and A_2;*

$$\begin{aligned}
&(X-\mu_1)^T A_1 (X-\mu_1) + (X-\mu_2)^T A_2 (X-\mu_2) \\
&\quad = (X-\mu^*)^T (A_1+A_2)(X-\mu^*) + (\mu_1-\mu_2)^T A_1 (A_1+A_2)^{-1} A_2 (\mu_1-\mu_2),
\end{aligned}$$

where $\mu^* = (A_1+A_2)^{-1}(A_1\mu_1 + A_2\mu_2)$.

In our context for the linear mixed model, we substitute γ_i for X above. We take $\mu_1 = \psi_i$ and $A_1 = \Sigma_i^{-1}$ for the first quadratic portion. For the second, we choose $\mu_2 = \hat{\gamma}_i = (Z_i^T Z_i)^{-1} Z_i^T (Y_i - X_i\beta)$ and $A_2 = \sigma^{-2} Z_i^T Z_i$. Using the complete-the-square formula, we have

$$\begin{aligned}
&(\gamma_i - \psi_i)^T \Sigma_i^{-1} (\gamma_i - \psi_i) + \frac{1}{\sigma^2} (\gamma_i - \hat{\gamma}_i)^T Z_i^T Z_i (\gamma_i - \hat{\gamma}_i) \\
&\quad = (\gamma_i - \gamma_i^*)^T \left(\Sigma_i^{-1} + \frac{1}{\sigma^2} Z_i^T Z_i\right) (\gamma_i - \gamma_i^*) \\
&\qquad + \frac{1}{\sigma^2} (\psi_i - \hat{\gamma}_i)^T \Sigma_i^{-1} \left(\Sigma_i^{-1} + \frac{1}{\sigma^2} Z_i^T Z_i\right)^{-1} Z_i^T Z_i (\psi_i - \hat{\gamma}_i),
\end{aligned}$$

where

$$\gamma_i^* = \left(\Sigma_i^{-1} + \frac{1}{\sigma^2} Z_i^T Z_i\right)^{-1} \left(\Sigma_i \psi_i + \frac{1}{\sigma^2} Z_i^T (Y_i - X_i \beta)\right).$$

We use this new expression to rewrite the joint density for Y_i and γ_i.

$$\begin{aligned}
f(Y_i, \gamma_i \mid \theta) &= (2\pi)^{-(n+q)/2} \sigma^{-n} |\Sigma_i|^{-1/2} \\
&\quad \times \exp\left(-\frac{1}{2\sigma^2}\left(\tilde{Y}_i - Z_i \hat{\gamma}_i\right)^T \left(\tilde{Y}_i - Z_i \hat{\gamma}_i\right)\right) \\
&\quad \times \exp\left(-\frac{1}{2\sigma^2}(\psi_i - \hat{\gamma}_i)^T \Sigma_i^{-1} \left(\Sigma_i^{-1} + \frac{1}{\sigma^2} Z_i^T Z_i\right)^{-1} Z_i^T Z_i (\psi_i - \hat{\gamma}_i)\right) \\
&\quad \times \exp\left(-\frac{1}{2}(\gamma_i - \gamma_i^*)^T \left(\Sigma_i^{-1} + \frac{1}{\sigma^2} Z_i^T Z_i\right)(\gamma_i - \gamma_i^*)\right)
\end{aligned}$$

Note that γ only appears in the final term, which has the form of a Normal kernel. This allows us to rewrite the joint density as follows:

$$\begin{aligned}
f(Y_i, \gamma_i \mid \theta) &= (2\pi\sigma^2)^{-n/2} |\Sigma_i|^{-1/2} \left|\Sigma_i^{-1} + \frac{1}{\sigma^2} Z_i^T Z_i\right|^{-1/2} \\
&\quad \times \exp\left(-\frac{1}{2\sigma^2}\left(\tilde{Y}_i - Z_i \hat{\gamma}_i\right)^T \left(\tilde{Y}_i - Z_i \hat{\gamma}_i\right)\right) \\
&\quad \times \exp\left(-\frac{1}{2\sigma^2}(\psi_i - \hat{\gamma}_i)^T \Sigma_i^{-1} \left(\Sigma_i^{-1} + \frac{1}{\sigma^2} Z_i^T Z_i\right)^{-1} Z_i^T Z_i (\psi_i - \hat{\gamma}_i)\right) \\
&\quad \times (2\pi)^{-q/2} \left|\Sigma_i^{-1} + \frac{1}{\sigma^2} Z_i^T Z_i\right|^{1/2} \exp\left(-\frac{1}{2}(\gamma_i - \gamma_i^*)^T \left(\Sigma_i^{-1} + \frac{1}{\sigma^2} Z_i^T Z_i\right)(\gamma_i - \gamma_i^*)\right) \\
&= f(Y_i \mid \theta) \times f(\gamma_i \mid Y_i, \theta)
\end{aligned}$$

Since observations on different clusters are assumed to be conditionally independent, the marginal for the entire data set is just $\prod_{i=1}^k f(Y_i \mid \theta)$, and we thus obtain a numerical approx-

imation to DIC_m using the marginal density:

$$
\begin{aligned}
f(Y \mid \theta) = (2\pi\sigma^2)^{-kn/2}|\Sigma|^{-1/2} \left|\Sigma^{-1} + \frac{1}{\sigma^2}\mathbf{Z}^T\mathbf{Z}\right|^{-1/2} \\
\times \exp\left(-\frac{1}{2\sigma^2}\left(\tilde{\mathbf{Y}} - \mathbf{Z}\hat{\boldsymbol{\gamma}}\right)^T\left(\tilde{\mathbf{Y}} - \mathbf{Z}\hat{\boldsymbol{\gamma}}\right)\right) \\
\times \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\psi} - \hat{\boldsymbol{\gamma}})^T\Sigma^{-1}\left(\Sigma^{-1} + \frac{1}{\sigma^2}Z^TZ\right)^{-1}Z^TZ(\boldsymbol{\psi} - \hat{\boldsymbol{\gamma}})\right),
\end{aligned}
$$

where $\hat{\boldsymbol{\gamma}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T(\mathbf{Y} - \mathbf{X}\beta)$.

We remind the reader that, given a sample from the joint posterior, $P(\theta, \gamma \mid y)$, the iterates for θ are from the marginal posterior $P(\theta \mid y)$, say $\{\theta^{(s)} : s = 1, 2, ...B\}$. Then

$$
\begin{aligned}
p_{Dm} &\doteq -\frac{2}{B}\sum_{s=1}^{B}\log f\left(\mathbf{Y} \mid \theta^{(s)}\right) + 2\log f\left(\mathbf{Y} \mid \hat{\theta}\right) \\
DIC_m &\doteq -\frac{4}{B}\sum_{s=1}^{B}\log f\left(\mathbf{Y} \mid \theta^{(s)}\right) + 2\log f\left(\mathbf{Y} \mid \hat{\theta}\right)
\end{aligned}
\tag{3.11}
$$

The MCMC sample can be obtained using OpenBUGS, JAGS, STAN or any other available sampling software.

In the next chapter, we will consider a special case of the generalized linear mixed model (GLMM), where closed-form marginalization is not possible, but where the expression we have here derived points the way toward a new method for approximate marginalization.

Chapter 4

Marginalization for DIC – Part II

In the previous chapter, we presented an overview of mixed modeling focusing on the simplest case: the linear mixed model. We discussed various DIC calculation methods and argued for a marginalized approach. We presented two methods for marginalizing linear mixed models in the Bayesian setting: (1) the preprocessing approach, which involves MCMC sampling based on the marginal likelihood $L(\theta \mid y) \propto \int f(y \mid \theta, \gamma) f(\gamma \mid \theta) d\gamma$ when an analytical form exists, and allows us to directly monitor the posterior deviance in packaged software; and (2) the postprocessing approach, which uses an analytical form for $L(\theta \mid y)$ or numerical methods to approximate the DIC based on existing samples for (θ, γ) that were not generated from the marginal distribution.

In this chapter, we expand the concept of marginalization to generalized linear mixed models—a wider class where the response data y are assumed to come from an exponential family, but not necessarily the normal family. We begin with a discussion of generalized linear mixed models (GLMMs) and an explanation of why the methods in the previous chapter won't work to marginalize this class. We then introduce a new method to efficiently approximate the marginal density for a special case of GLMMs. This special case—which

will call "repeated exchangeable observations (REO) GLMMs"—arises when response data within each cluster can be permuted without changing the resulting inferences. Chapter 5 shows how to extend the method to a broader class of GLMMs.

4.1 Generalized Linear Mixed Models

This section provides a short introduction to generalized linear mixed models. We review their history and provide a technical definition. We then discuss the two contexts—overdispersion and correlation—that most frequently prompt their use. Finally, we present the logistic (binomial) and log-linear (Poisson) GLMM models with normal random effects, which we use extensively throughout the chapter.

4.1.1 Development

Nelder and Wedderburn (1972) proposed the framework for the generalized linear model (GLM), in which a linear component $X\beta$ is equated, through some invertible function $q(\cdot)$, with the mean of the response variable $\mathbf{Y}$. We repeat our notation from Section 3.2.1, with i serving as a cluster index and j serving as an observation index. Then

$$\mathrm{E}\big[y_{ij}\big] = q^{-1}(X_{ij}\beta) \qquad \text{or equivalently} \qquad q(\mathrm{E}\big[y_{ij}\big]) = X_{ij}\beta.$$

Work in the early 1980s began to incorporate random effects into this type of modeling (e.g. Williams 1982; Breslow 1984), primarily as a way to handle overdispersion in generalized linear models. This culminated in the development of the GLMM as an extension of the GLM structure.

In general, a GLMM is a model of the form:

$$\begin{aligned} y_{ij} \mid X_{ij}, \gamma_i &\overset{\text{iid}}{\sim} \text{exponential family} \\ \mathrm{E}\big[y_{ij}\big] &= \mu_{ij} \\ q(\mu_{ij}) &= X_{ij}\beta + Z_{ij}\gamma_i \\ \gamma_i &\overset{\text{iid}}{\sim} f(\theta_0) \end{aligned} \tag{4.1}$$

Here y_{ij}, X_{ij}, β, and $q(\cdot)$ are defined as in the GLM, except that now we add γ_i, a vector of random effects that is specific to cluster i with some distribution that depends on parameter vector θ_0. Z_{ij} specifies a row from the $\mathbf{Z}_i$ design matrix for random effects on cluster i, and $\mathbf{Z}_i$ is often a submatrix of $\mathbf{X}_i$—that is, each column of $\mathbf{Z}_i$ is also a column of $\mathbf{X}_i$. We refer the reader to Section 3.2.1 if more notational detail is needed.

We assume a standard multivariate normal distribution for the random effects, $\gamma_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \Sigma)$, with $\theta_0 = \Sigma$. Other common distributions used for γ_i involve non-zero-mean normals and mixtures of normals. One particular alternative deserves special note.

When Z_{ij} is a submatrix of X_{ij}, a centering parametrization can be used, and can often improve MC sampling efficiency. To understand how this alternative parametrization is structured, assume X_{ij} can be partitioned into two sets of columns as

$$X_{ij} = \begin{bmatrix} X_{ij}^{(1)} & X_{ij}^{(2)} \end{bmatrix},$$

with the submatrix $Z_{ij} = X_{ij}^{(1)}$. Assume, further, that we partition β as

$$\beta = \begin{bmatrix} \beta^{(1)} \\ \\ \beta^{(2)} \end{bmatrix}$$

so that

$$X_{ij}\beta = \begin{bmatrix} X_{ij}^{(1)} & X_{ij}^{(2)} \end{bmatrix} \begin{bmatrix} \beta^{(1)} \\ \beta^{(2)} \end{bmatrix}.$$

Then let $\xi_i \overset{\text{iid}}{\sim} \text{N}(0, \Sigma)$ be the random effects under the standard parametrization. We write

$$\begin{aligned} q(\mu_{ij}) &= X_{ij}\beta + Z_{ij}\xi_i \\ &= X_{ij}^{(1)}\beta^{(1)} + X_{ij}^{(2)}\beta^{(2)} + Z_{ij}\xi_i \\ &= Z_{ij}\beta^{(1)} + +X_{ij}^{(2)}\beta^{(2)} + Z_{ij}\xi_i \\ &= X_{ij}^{(2)}\beta^{(2)} + Z_{ij}\left(\beta^{(1)} + \xi_i\right) \\ &\equiv X_{ij}^{(2)}\beta^{(2)} + Z_{ij}\gamma_i. \end{aligned}$$

Then $\gamma_i \overset{\text{iid}}{\sim} \text{N}\left(\beta^{(1)}, \Sigma\right)$, and $\theta_0 = \{\beta^{(1)}, \Sigma\}$. This is the aforementioned centering parametrization for the GLMM. A. E. Gelfand et al. (1996) demonstrate that this parametrization often improves convergence in the GLMM setting.

4.1.2 Use

One use for GLMMs, as mentioned above, is with exponential-family response data that exhibit overdispersion. The most common forms of GLMM—the binomial and Poisson—both involve distributions that exhibit a mean-variance relationship. Even when these distributions for the data might seem like a natural fit, it is not uncommon to find that the data exhibit more variability than would be expected.

The other primary use for GLMMs is to provide correlation structure for repeated response data and/or clustered data. Modeling correlation is generally required for data with a hierarchical structure—e.g. when there are multiple observations within clusters, and where observations taken on the same cluster are expected to share similarities not fully captured

by knowing the model covariates. This is the same philosophical context in which linear mixed models are useful, but now applied to a wider range of possible response data.

4.1.3 The Binomial and Poisson GLMMs

Throughout this chapter and the next, we focus attention on the binomial (logistic) and Poisson (log-linear) GLMMs. The terms logistic and log-linear stem from the choice of link function, $q(\cdot)$, used for response data of the corresponding type[1]. These are the most commonly used GLMMs, so we pay special attention to them throughout the dissertation—both by providing specific methodological details and by focusing on them in simulations. We do not provide details for alternative binomial link functions such as the probit and clog-log, but adapting the general form of the methods we develop to these link functions should be straightforward.

Here we present standard GLMMs for the logistic and log-linear links. These have been written in sufficient generality to (1) allow for either a standard or a centering parametrization through ψ_i and (2) deal with observation-specific differences in binomial trials or Poisson interval sizes through m_{ij}. We discuss the use of ψ_i below, after presenting the models.

Binomial (logistic):

$$\begin{aligned} Y_{ij} &\perp\!\!\!\perp \text{Bin}\left(m_{ij}, \pi_{ij}\right) \\ q\left(\pi_{ij}\right) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) &= X_{ij}\beta + Z_{ij}\gamma_i \\ \gamma_i &\perp\!\!\!\perp N(\psi_i, \Sigma) \end{aligned} \tag{4.2}$$

[1]Technically, the link function $q(\cdot)$ for the logistic model is known as the logit function; and the logistic function, also known as the expit function, is its inverse. The use of 'logistic' as the name for the model is, nonetheless, standard practice.

Poisson (log-linear):

$$\begin{aligned} Y_{ij} &\perp\!\!\!\perp \text{Pois}\left(m_{ij}\lambda_{ij}\right) \\ q\left(\lambda_{ij}\right) = \log\left(\lambda_{ij}\right) &= X_{ij}\beta + Z_{ij}\gamma_i \\ \gamma_i &\perp\!\!\!\perp N(\psi_i, \Sigma) \end{aligned} \tag{4.3}$$

Here and elsewhere in the dissertation, we write our random effects as $\gamma_i \perp\!\!\!\perp N(\psi_i, \Sigma)$. This is done to ensure that expressions and methods can be used with both a standard parametrization and a centering parametrization (c.f. Section 4.1.1) for the random effects. When a standard parametrization is used, $\psi_i \equiv 0$ for all i, and no further discussion is necessary.

If a centering parametrization is used, however, we must be cognizant of identifiability issues. For example, let β_1 be a grand mean term, present in the model for each datapoint y_{ij}; and let ψ_i be equal to a non-zero constant for all i. Then only the sum $\beta_1 + \psi$ is identifiable– the components β_1 and ψ themselves are not identifiable. Our description in Section 4.1.1 indicates how the model is constructed if a centering parametrization is to be implemented. Partition the fixed effects design matrix, $\mathbf{X}$, into: $\mathbf{X}^{(1)}$, those columns shared by the random effects design matrices, $\mathbf{Z}_i$, $i \in \{1, ..., k\}$; and $\mathbf{X}^{(2)}$, those columns not shared by the $\mathbf{Z}_i$'s. Partition β similarly into $\beta^{(1)}$ and $\beta^{(2)}$. Then the centering parametrization uses $\psi_i = \beta^{(1)}$. Note also that

$$\mathbf{X}\beta = \mathbf{X}^{(1)}\beta^{(1)} + \mathbf{X}^{(2)}\beta^{(2)}.$$

Then $\mathbf{X}^{(1)}\beta^{(1)}$ is removed from the fixed effects portion of the linear model to avoid non-identifiability. Therefore, when using the centering parametrization, $X_{ij}\beta$ in the above expressions must be replaced with $X_{ij}^{(2)}\beta^{(2)}$, since $Z_{ij}\gamma_i$ will already contain the redundant part of the model.

4.2 Marginalization in GLMMs

Our goal with GLMMs is to obtain a numerical approximation to the DIC_m for use in model selection. Following our discussion in the preceding chapter, we believe model selection in mixed modeling should use DIC_m. Unfortunately, our methods for obtaining DIC_m in the LMM setting do not extend to the GLMM, as we now discuss. We will also discuss how the available tools for tackling this problem in the GLMM setting are computationally expensive and become intractable when more than two random terms are used. In the following section we explain: the difficulties we encounter in seeking to marginalize GLMMs, techniques that are currently available for GLMM marginalization, and related work.

4.2.1 Preprocessing Won't Work

As we discussed in Section 3.4.1.1, the goal of preprocessing marginalization is to conduct MCMC sampling that is based on an analytically tractable marginal likelihood for θ, i.e. $L(\theta \mid y) \propto \int f(y \mid \gamma, \theta) p(\gamma \mid \theta) d\gamma$. The deviance, $D(\theta)$, is monitored for posterior inference.

In the linear mixed model, it is possible to write the marginal density for the data conditional on the parameters in closed form. This results in Equation (3.7), a multivariate normal model for the data vector $\mathbf{Y}$ with a mean given by the linear model and a structured covariance matrix. No such closed form solution exists for marginalization in the GLMM setting.

The preprocessing approach will not work. We thus consider options for postprocessing—approximation methods that can be used when a posterior sample is available.

4.2.2 Postprocessing Is Difficult

In this subsection, we consider the postprocessing approach to marginalizing GLMMs. Our goal, as in Section 3.4.1.2, is to sample from the full model involving γ, and to obtain an approximate expression for the marginal density. We use posterior iterates in connection with this expression to obtain a numerical approximation to DIC_m. Below, we discuss how this can be done with two numerical integration methods: MC integration and Gaussian quadrature.

A significant advantage to the postprocessing approach is that our expression for the marginal density need not be analytic. As we stated above, no analytic expression exists for marginalizing GLMMs. If there were, we would be able to use the simple form of the marginal DIC formulae from Section 3.3.1.3:

$$\begin{aligned} p_{Dm} &= \mathrm{E}_{\theta|y}[-2\log L(\theta \mid y)] + 2\log L(\hat{\theta} \mid y), \\ DIC_m &= 2\,\mathrm{E}_{\theta|y}[-2\log L(\theta \mid y)] + 2\log L(\hat{\theta} \mid y). \end{aligned}$$

Since there is no closed form marginal likelihood, we instead consider the following formulae:

$$\begin{aligned} p_{Dm} &= \mathrm{E}_{\theta|y}\Big[-2\log \mathrm{E}_{\gamma|\theta}[L(\theta,\gamma \mid y)]\Big] + 2\log \mathrm{E}_{\gamma|\theta}\Big[L(\hat{\theta},\gamma \mid y)\Big] \\ &= \mathrm{E}_{\theta|y}\left[-2\log \int L(\theta \mid y,\gamma)d\gamma\right] + 2\log \int L(\hat{\theta} \mid y,\gamma)d\gamma \\ DIC_m &= 2\,\mathrm{E}_{\theta|y}\Big[-2\log \mathrm{E}_{\gamma|\theta}[L(\theta,\gamma \mid y)]\Big] + 2\log \mathrm{E}_{\gamma|\theta}\Big[L(\hat{\theta},\gamma \mid y)\Big] \\ &= 2\,\mathrm{E}_{\theta|y}\left[-2\log \int L(\theta \mid y,\gamma)d\gamma\right] + 2\log \int L(\hat{\theta} \mid y,\gamma)d\gamma \end{aligned}$$

where $\hat{\theta} = \mathrm{E}_{\theta,\gamma|y}[\theta]$ as in Section 3.3.1.3.

These expressions can be approximated by numerical integration methods—though such approximation is computationally expensive. The most natural approaches involve MC integration and Gaussian quadrature. We provide two expressions for p_{Dm} and DIC_m, one based on MC integration and the other based on Gaussian quadrature.

Monte Carlo integration (Hammersley and Handscomb 1964) over γ begins by taking an MC sample of size B_t from the joint posterior for θ and γ: $\{(\theta^{(1)}, \gamma^{(1)}), ..., (\theta^{(B_t)}, \gamma^{(B_t)})\}$. Then for each $\theta^{(t)}$ and for $\hat{\theta}$, we draw an additional MC sample of size B_g from the conditional posterior $p(\gamma \mid y, \theta^{(t)})$. We denote these additional γ's as $\gamma_{(t)}^{(s)} \in \{\gamma_{(t)}^{(1)}, ..., \gamma_{(t)}^{(B_g)}\}$, where the (t) subscript denotes the posterior iterate $\theta^{(t)}$ that γ is conditioned on, and the (s) superscript denotes that this is the s^{th} of B_g observations on that conditional posterior.

Then we can numerically approximate p_{Dm} and DIC_m as:

$$p_{Dm} \doteq -\frac{2}{B_t}\sum_{t=1}^{B_t}\log\left(\frac{1}{B_g}\sum_{s=1}^{B_g} f\left(y \mid \gamma_{(t)}^{(s)}, \theta^{(t)}\right)\right) + 2\log\left(\frac{1}{B_g}\sum_{s=1}^{B_g} f\left(y \mid \gamma_{\hat{\theta}}^{(s)}, \hat{\theta}\right)\right)$$

$$DIC_m \doteq -\frac{4}{B_t}\sum_{t=1}^{B_t}\log\left(\frac{1}{B_g}\sum_{s=1}^{B_g} f\left(y \mid \gamma_{(t)}^{(s)}, \theta^{(t)}\right)\right) + 2\log\left(\frac{1}{B_g}\sum_{s=1}^{B_g} f\left(y \mid \gamma_{\hat{\theta}}^{(s)}, \hat{\theta}\right)\right)$$

This is a costly solution to the marginalization problem. We must produce $(B_t+1)B_g$ separate MCMC samples in order to approximate DIC_m. The time it takes to run a chain for a GLMM is non-ignorable, and the MC integration approach to marginalization effectively multiplies that time by B_t. With $B_t = 5000$ and 1s MCMC processing time for a model, marginalization by MC integration takes about an hour and a half. With $B_t = 5000$ and 60s MCMC processing time, this approach takes around 3.5 days. Some of the simulations we conducted for this work involved MCMC chains that took more than two days to run; MC integration for these simulations could potentially take longer than the professional lifetime of the author.

A more appealing solution is offered by Gaussian quadrature (Smith et al. 1985). Here, instead of sampling from $\gamma \mid y, \theta^{(t)}$, we evaluate $L(\theta^{(t)} \mid y, \gamma) \propto f(y, \gamma \mid \theta^{(t)})$ over a pre-specified grid of γ values, weight the resulting values according to pre-specified weights, and sum them to obtain a numerical approximation to the marginal likelihood $L(\theta^{(i)} \mid y)$. Approximating DIC_m still requires $(B_t + 1)B_n$ evaluations of the likelihood function, where B_n is the number of quadrature nodes in the grid over which the function is evaluated. This is similar to the $(B_t + 1)B_g$ evaluations necessary for MC integration except that $B_n \ll B_g$ in general and there is no additional computationally expensive simulation involved with the quadrature approach.

Mathematically, let $u^{(s)}$, $s \in \{1, ..., B_n\}$ be the pre-specified grid of values at which our function will be evaluated. Let $v^{(s)}$, $s \in \{1, ..., B_n\}$ be the pre-specified weights associated with those nodes. Then for a function, $f(x)$, the idea is that

$$\int f(x)dx = \sum_{s=1}^{B_n} f(u^{(s)})v^{(s)}.$$

In truth, the method is a little more complicated, as we explain below.

Quadrature usually necessitates a change of variables in the joint density, $f(y, \gamma \mid \theta)$. The most common quadrature method, Gauss-Legendre, requires that the integrated variable live on the interval $(-1, 1)$. Consider the case where the dimensionality of γ_i, q, is 1. We adopt our formal mixed modeling notation (Section 3.2.1) here, to make it clearer how quadrature works the mixed modeling setting. Since we assume γ_i to be univariate normal, this requires a transformation in scale. A modified logistic transformation works well, where

$$\nu_i = 2\left(\frac{\exp(\gamma_i)}{1 + \exp(\gamma_i)} - \frac{1}{2}\right), \qquad \gamma_i = \log\left(\frac{1 + \nu_i}{1 - \nu_i}\right).$$

Then our joint density for an individual i, in (Y_i, ν_i) as opposed to (Y_i, γ_i), is given by

$$f_{\nu_i}(Y_i, \nu_i \mid \theta) = f_{\gamma_i}\left(Y_i, \log\left(\frac{1+\nu_i}{1-\nu_i}\right) \mid \theta\right) \frac{2}{1-\nu_i^2}$$

and

$$L(\theta \mid \mathbf{Y}, \boldsymbol{\nu}) \propto \prod f_{\nu}(Y_i, \nu_i \mid \theta).$$

Formally, let $u_L^{(s)}$ and $v_L^{(s)}$ be defined as above, using the pre-specified nodes and weights for Gauss-Legendre quadrature, which are based on Legendre polynomials. Then the Gauss-Legendre quadrature approximation to $f(\mathbf{Y} \mid \theta)$ is given by

$$\begin{aligned} f(\mathbf{Y} \mid \theta) &= \prod_{i=1}^{k} f(Y_i \mid \theta) \\ &= \prod_{i=1}^{k} \int f_{\nu_i}(Y_i, \nu_i \mid \theta) d\nu_i \\ &= \prod_{i=1}^{k} \int f_{\gamma_i}(Y_i, \frac{\frac{\nu_i+1}{2}}{1-\frac{\nu_i+1}{2}} \mid \theta) \frac{2}{1-\nu_i^2} d\nu_i \\ &\doteq \prod_{i=1}^{k} \left(\sum_{s=1}^{B_n} f_{\gamma_i}\left(Y_i, \frac{\frac{u_L^{(s)}+1}{2}}{1-\frac{u_L^{(s)}+1}{2}} \mid \theta\right)\right) \frac{2}{1-(u_L^{(s)})^2} v_L^{(s)} \end{aligned} \tag{4.4}$$

Note that Gaussian quadrature approximates univariate integration. Multivariate integration is possible by creating a multidimensional grid of nodes and performing nested quadrature steps on each dimension. Thus, if the solution to a three-dimensional integral is required and 20 nodes are used for evaluation in each dimension, the resulting procedure considers the function at $20^3 = 8000$ unique points. Fortunately, because of the independence between clusters we assume in the above example, we do not have to use multivariate quadrature to deal with the clusters simultaneously. The multivariate integral over $\boldsymbol{\gamma}$ is equal to the product of the individual integrals over the γ_i's. However if q, the dimensionality of γ_i, increases, computing time will also increase exponentially.

Other Gaussian quadrature methods, most notably Gauss-Hermite, require less reformulation of the joint density. Gauss-Hermite quadrature approximates integrals of the kind $\int_{-\infty}^{\infty} e^{-x^2} f(x)dx$, which is naturally appealing since $f(Y_i, \gamma_i \mid \theta)$ includes a normal kernel for γ_i. Details of the Gauss-Hermite method are similar to the Gauss-Legendre method above. They are provided in detail in the appendix—both to let the reader see how quadrature works for multivariate γ_i and because we will make use of the Gauss-Hermite method in a small number of simulations.

Consideration of alternative quadrature methods, however, leads to one of the significant complications with these approaches: deciding on the number of nodes, B_n. Quadrature rules are designed to give exact results for polynomials of order $2B_n - 1$ or less. Because every continuous function can be approximated by a polynomial of sufficiently high order, we know the quadrature approach gives us good results when B_n is large enough—but how large is large enough? The answer to this question depends on the type of quadrature method (e.g. Gauss-Legendre, Gauss-Hermite), as well as the characteristics of the likelihood function itself. The order of polynomial required to sufficiently approximate a function will depend on its characteristics such as its peakedness; and since the function employs different transformations depending on the quadrature method chosen, these characteristics will depend on that choice of method.

This means that an accurate numerical approximation to the integral depends on choosing a grid that includes enough points to capture behavior near the peak. Although not very computationally efficient, perhaps the best way to determine whether enough nodes have been used is to perform repeated quadrature approximations at an increasing number of nodes, observing when the results appear to converge to a stable answer to the numerical integral.

4.2.3 Other Approaches Involving Hierarchical Models

A frequentist method for longitudinal GLM data is the generalized estimating equation approach (GEE; Liang and Zeger 1986). The GEE approach provides consistent estimates of the regression parameters, β, and their covariance structure in the scenario considered in this chapter. The GEE approach only provides estimates of β and its covariance, and inferences rely on asymptotic normality. We are thus not able to use this approach for making posterior inferences— which do not rely on asymptotics, nor are they limited to point and interval estimates of regression coefficients.

Work on model selection for GLMMs, beyond what we have already discussed, has largely been directed at different model selection criteria and selection paradigms—see, e.g. Saefken et al. (2014) for work on the conditional AIC in GLMMs; or Overstall and Forster (2010) and Sinharay and Stern (2005) for work on the Bayes factor in GLMMs. When these methods deal with marginalization relative to the random effects, Sinharay and Stern suggest numerical integration as discussed above, or importance sampling methods for high-dimensional γ_i. Although they do not specify what constitutes high dimensionality, based on our own work we believe the numerical integration methods we describe should be avoided for $\dim(\gamma_i) \geq 3$.

Cai and Dunson (2006) provide a marginalization formula for GLMMs based on a second-order Taylor expansion of $L(\theta|y,\gamma)$ around $\mathrm{E}_{\gamma|\theta}[\gamma]$. This is done to provide a Bayesian variable selection procedure based on the use of mixture priors (e.g. spike-and-slab). We consider this approach of special note, because it is the closest method we found to our own—which also uses a Taylor expansion involving the random effects as one step in approximating a marginal density for GLMMs. However, our method takes advantage of having random effects that are modeled with a multivariate normal distribution to yield a notably simpler approximation. We discuss this method after we have developed our method in Section 4.3.1.

4.3 A Limited Marginalization Approach

Our goal in this chapter is to develop a computationally efficient method for approximating DIC_m. Posterior samples provided by most Bayesian analysis programs make the approximation to DIC_j straightforward, but approximating DIC_m requires either additional numerical integration or some form of analytic approximation.

The simplest way to obtain DIC_m would be to adopt the quadrature approach: applying Equation (4.4) to each element of the posterior sample $\theta^{(s)}$, $s \in \{1, ..., B_t\}$, and to the posterior mean[2] $\overline{\theta} = \frac{1}{B_t} \sum_s \theta^{(s)}$, and then substituting the resulting values for $f(\boldsymbol{Y} \mid \theta)$ into Equation (3.11) to obtain the approximation to DIC_m. As discussed above, this can create an enormous computing burden. A good marginalization approach will involve a reduced computational burden so that approximation can be done quickly and accurately.

Further, one of the central appeals of DIC as a Bayesian model selection criterion is the ease with which it can be calculated from an MCMC sample (SBCV; R. R. Christensen, Johnson, et al. 2010; Gelman et al. 2013). We believe a good marginalization approach will retain this feature: it should be calculable based on existing MCMC output, rather than requiring it be built into the MCMC procedure itself.

4.3.1 Approximate Marginalization through Taylor Expansion

We begin this section by reviewing Taylor's method for function approximation. We then outline our methodological development to highlight key points for the reader, and discuss the REO special case of GLMMs. We follow this with a formal statement of our method, which proceeds from the application of Taylor's theorem, Proposition 3.1, and a discussion

[2]Or whatever point estimate of θ is preferred; though like SBCV, we focus our attention on the posterior mean of θ.

of approximation error under our method. Finally, we conclude this section by giving the functional form of our approximation in both binomial and Poisson GLMMs.

4.3.1.1 Taylor expansion

Taylor's theorem is a method for functional approximation that dates back to the invention of calculus. It is named for Brook Taylor who derived it in 1712—although it was known by others as early as 1670, and explicit definition of the approximation error would not be specified until Lagrange in 1772 (Kline 1972). We refer to the formulation given by R. R. Christensen (1997), which states that for some function $g(\cdot)$ with at least second-order derivatives, and with x and a in the domain of g,

$$g(x) = g(a) + \dot{g}(a)\,(x-a) + \frac{1}{2}\,(x-a)^T\,\ddot{g}(a)\,(x-a) + r(g,x,a), \tag{4.5}$$

where

$$\dot{g}(x) \equiv \frac{\partial}{\partial x} g(x) \qquad \text{and} \qquad \ddot{g}(x) \equiv \frac{\partial^2}{\partial x \partial x^T} g(x).$$

We use dot notation for derivatives commonly appearing throughout this work. In this chapter, where we will only be concerned with univariate derivatives, we also use $g^{(3)}(\cdot)$ for the 3^{rd} derivative of $g(\cdot)$. We occasionally use ${}^{d}/_{dx}$ notation as well when we think the additional clarity will help the reader.

The function $r(g,x,a)$ here is the remainder for the second-order Taylor approximation, which is of order $o(||x-a||^2)$, meaning that

$$\forall \varepsilon \exists \delta \text{ s.t. } ||x-a|| < \delta \Rightarrow |r(g,x,a)| \leq \varepsilon (x-a)^T (x-a).$$

We also consider the Lagrange form of the second-order remainder when the third derivative exists,

$$r(g, x, a) = \frac{g^{(3)}(\xi_L)}{3!}(x - a)^3, \tag{4.6}$$

where ξ_L is some real number between x and a. This is a mean-value form for the Taylor remainder, and can help us gain insight into the behavior of the approximation.

In the following sections, we provide a novel application of this formula for GLMM models when random effects are assumed to have a normal distribution, and we show how this can be used to obtain a new approximate marginal form.

4.3.1.2 An outline of our development

The marginalization approach that we develop below is designed for use with a subset of GLMMs where each cluster is associated with multiple observations and all covariates are constant within clusters—that is, random intercept models with repeated exchangeable observations (REO) on each individual. Chapter 5 details a more general marginalization approach that extends this work and can be used for all GLMMs.

Our marginalization approach makes use of the exponential family form of GLMM models. This lets us write a joint density for data y and normally-distributed random effects γ where γ's are isolated inside an exponential function. The conditional density of y given γ and θ is

$$f(y \mid \gamma, \theta) = h^*(y) \exp\left(g^*(y \mid \gamma, \theta)\right),$$

where $g^*(y \mid \gamma, \theta)$ is the exponential component of the pdf (or pmf) for y. Recall that our concern in marginalization focuses on the integral

$$\int L(\theta \mid y, \gamma) d\gamma = \mathrm{E}_{\gamma|\theta}[L(\theta, \gamma \mid y)] = \int f(y \mid \gamma, \theta) p(\gamma \mid \theta) d\gamma.$$

We then focus on the interpretation of $f(y \mid \gamma, \theta)$ as a function of γ for fixed θ. This allows us to write

$$\int L(\theta \mid y, \gamma) d\gamma = h(y, \theta) \int \exp\left(g(\gamma \mid y, \theta)\right) p(\gamma \mid \theta) d\gamma,$$

which is possible because $f(y \mid \gamma, \theta)$ depends on γ only through the exponent. For brevity, we use $g(\gamma) \equiv g(\gamma \mid y, \theta)$, except where the longer form is necessary for the reader's understanding.

We use a second-order Taylor approximation for $g(\gamma)$ centered on some $\hat{\gamma}$. In the instance described, where multiple observations are obtained for each individual, and with a single set of individual level covariates that correspond to all of their observations, it turns out that $\hat{\gamma}$ can be taken to be the average across repeated observations. Crucially, this average is the solution to $\dot{g}(\hat{\gamma}) = 0$. As we will see, the Taylor expansion involves a quadratic form in γ, which can then combined with the quadratic form in the model for $\gamma \mid \theta$ by using the complete-the-square formula (Proposition 3.1). The resulting approximation makes it easy to analytically integrate γ_i out of the approximation to $f(Y_i, \gamma_i \mid \theta)$.

Our objective is to approximate, along the lines just discussed,

$$\begin{aligned} L(\theta \mid y) &\propto f(y \mid \theta) \\ &= \int f(y, \gamma \mid \theta) d\gamma \\ &= \int f(y \mid \gamma, \theta) p(\gamma \mid \theta) d\gamma. \end{aligned}$$

If we could obtain a closed form expression for $f(y, \gamma \mid \theta) = f(\gamma \mid y, \theta) f(y \mid \theta)$, as in the LMM case, the problem would be immediately solved. This, then, is our strategy: to find an approximation to $f(y, \gamma \mid \theta)$ that permits this factorization.

4.3.1.3 Repeated exchangeable observations (REO)

In the previous section, we mentioned that the method we develop here is for random intercept models with repeated exchangeable observations (REO) on each individual/cluster. We will discuss the nature of the REO special case further, focusing on the differences between REO and non-REO settings, in Section 5.1.1.

Having repeated exchangeable observations allows us to "estimate" the random effects using within-cluster averages of the response data. The ability to do this simplifies our method, which is why we first deal with this special case before moving onto a more general method in the next chapter.

In the setting described, we have $X_i = \begin{bmatrix} x_{i1} & \dots & x_{ip} \end{bmatrix} \otimes J_n$ and $\mathbf{Z} = J_n \otimes I_k$. That is, covariates x vary between clusters but not within them, and γ_i's are a constant random effect for each cluster. Then

$$X_i \beta + Z_i \gamma_i = \left(\begin{bmatrix} x_{i1} & \dots & x_{ip} \end{bmatrix} \beta + \gamma_i \right) J_n,$$

and we can write

$$\overline{Y}_i = \frac{1}{n} Y_i^T J_n = q^{-1} \left(\hat{\gamma}_i + \sum_{t=1}^{p} x_{it} \beta_t \right).$$

From here, we solve for $\hat{\gamma}_i$, which is a function of Y_i.

$$\hat{\gamma}_i(Y_i) = q(\overline{Y}_i) - \sum_{t=1}^{p} x_{it} \beta_t. \tag{4.7}$$

This choice of $\hat{\gamma}_i(Y_i)$ maximizes $f(Y_i \mid \gamma_i, \theta)$ in γ_i, and thus maximizes $g(\gamma_i)$. For brevity, we use $\hat{\gamma}_i$ instead of $\hat{\gamma}_i(Y_i)$ through most of this text, except in places where it is important to

the reader's understanding to explicitly recognize that $\hat{\gamma}_i$ is a function of Y_i (e.g. Proposition 4.1).

Note that as the number of observations per cluster, n, increases, $\hat{\gamma}_i \xrightarrow{\text{a.s.}} \gamma_i$. Since our observations are conditionally independent within clusters, the Strong Law of Large Numbers applies and we have $\overline{Y_i} \xrightarrow{\text{a.s.}} q^{-1}\left(\gamma_i + \sum_t x_{it}\beta_t\right)$. Using Slutsky's Theorem, if we assume that β is known, then

$$\begin{aligned} q(\overline{Y}_i) &\xrightarrow{\text{a.s.}} \gamma_i + \sum_{t=1}^{p} x_{it}\beta_t \qquad \Longleftrightarrow \\ q(\overline{Y}_i) - \sum_{t=1}^{p} x_{it}\beta_t &\xrightarrow{\text{a.s.}} \gamma_i \qquad \Longleftrightarrow \\ \hat{\gamma}_i &\xrightarrow{\text{a.s.}} \gamma_i. \end{aligned}$$

Applying Equation (4.5), the Taylor expansion of $g(\gamma_i)$ around $\hat{\gamma}_i$ gives

$$g(\gamma_i) = g(\hat{\gamma}_i) + \dot{g}(\hat{\gamma}_i)(\gamma_i - \hat{\gamma}_i) + \frac{1}{2}\ddot{g}(\hat{\gamma}_i)(\gamma_i - \hat{\gamma}_i)^2 + r(g, \gamma, \hat{\gamma}).$$

But because $\hat{\gamma}_i$ maximizes $g(\gamma_i)$, we know $\dot{g}(\hat{\gamma}_i) = 0$ and the above expression simplifies to

$$g(\gamma_i) = g(\hat{\gamma}_i) + \frac{1}{2}\ddot{g}(\hat{\gamma}_i)(\gamma_i - \hat{\gamma}_i)^2 + r(g, \gamma_i, \hat{\gamma}_i). \tag{4.8}$$

This analytic removal of $\dot{g}(\hat{\gamma}_i)(\gamma_i - \hat{\gamma}_i)$ is what simplifies the REO GLMM special case.

Further note that $\ddot{g}(\gamma_i)$ is the same as $\frac{\partial^2}{\partial \gamma_i^2}\,\ell(\gamma_i, \theta \mid Y_i)$, where $\ell(\gamma_i, \theta \mid Y_i) \propto \log f(Y_i \mid \gamma_i, \theta)$. This relates our approximation to the Fisher information for γ_i contained in Y_i, with known θ.

4.3.1.4 Approximate joint and marginal densities

Now that we have outlined the important elements of our method, we proceed to formally develop our marginal approximation for the REO GLMM setting.

Substituting Equation (4.7) into $f(y, \gamma \mid \theta)$ gives us an approximation to the joint density. Formally, following the definition of the GLMM used in Section 4.1 with $\Sigma = [\sigma^2]$, we have

$$\begin{aligned} f(Y_i, \gamma_i \mid \theta) &= h(Y_i, \theta) \exp\left(g(\hat{\gamma}_i) - \frac{1}{2}\left[-\ddot{g}(\hat{\gamma}_i)\left(\gamma_i - \hat{\gamma}_i\right)^2 \right] + r(g, \gamma_i, \hat{\gamma}_i) \right) \\ &\quad \times \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{\gamma_i - \psi_i}{\sigma} \right)^2 \right) \\ &= \frac{h(Y_i, \theta)}{\sqrt{2\pi\sigma^2}} \exp\left(g(\hat{\gamma}_i) - \frac{1}{2}\left[-\ddot{g}(\hat{\gamma}_i)\left(\gamma_i - \hat{\gamma}_i\right)^2 + \frac{1}{\sigma^2}\left(\gamma_i - \psi_i\right)^2 \right] + r(g, \gamma_i, \hat{\gamma}_i) \right) \\ &= \hat{f}(Y_i, \gamma_i \mid \theta) \exp\left(r(g, \gamma_i, \hat{\gamma}_i) \right). \end{aligned} \tag{4.9}$$

where $\hat{f}(Y_i, \gamma_i \mid \theta)$ is our approximate joint density for the data and the random effects, if we drop the remainder term.

Proposition 4.1. *Assume γ_i is q dimensional. Then if $\ddot{g}(\cdot)$ is continuous on a closed ball around γ_i,*

$$\hat{f}(Y_i, \gamma_i \mid \theta) - f(Y_i, \gamma_i \mid \theta) \xrightarrow{\text{a.s.}} 0 \qquad n \to \infty.$$

Proof. By assumption, there $\exists\, \delta$ such that $\ddot{g}(t)$ is continuous over the closed ball $B_\delta = \{t : ||t - \gamma_i|| \leq \delta\}$.

Since $\hat{\gamma}_i \xrightarrow{\text{a.s.}} \gamma_i$, $\Pr\left[\omega : \lim_{n\to\infty} \hat{\gamma}_i(\omega) = \gamma_i\right] = 1$. Denote as Ω the set of ω where this happens. Then $\forall\, \omega \in \Omega$, $\exists\, N(\omega)$ such that $n \geq N(\omega) \implies ||\hat{\gamma}_i(\omega) - \gamma_i|| \leq \delta$.

According to *A Course in Large Sample Theory* by Ferguson (1996), if $\ddot{g}(t)$ is continuous over B_δ, then Taylor's theorem indicates that, for $||s|| < \delta$,

$$g(\hat{\gamma}_i(\omega) + s) = g(\hat{\gamma}_i(\omega)) + \dot{g}(\hat{\gamma}_i(\omega))s + s^T \left(\int_0^1 \int_0^1 v\ddot{g}\left(\hat{\gamma}_i(\omega) + uvs\right) dudv \right) s.$$

Let $s = s(\omega) = \gamma_i - \hat{\gamma}_i(\omega)$ for $\omega \in \Omega$. We thus have

$$g(\gamma_i) - g(\hat{\gamma}_i(\omega)) = s(\omega)^T \left(\int_0^1 \int_0^1 v\ddot{g}\left(\hat{\gamma}_i(\omega) + uvs(\omega)\right) dudv \right) s(\omega),$$

since $\dot{g}(\hat{\gamma}_i(\omega)) = 0$ by construction.

Then for this choice of s, and with $n > N(\omega)$, we must have $||s(\omega)|| < \delta$. We know $||\ddot{g}(t)||$ is bounded (in each component of the $q \times q$ matrix) on the closed ball B_δ, because it is continuous on that closed ball. Then since $||\hat{\gamma}_i(\omega) + uvs(\omega) - \hat{\gamma}_i(\omega)|| = ||uvs(\omega)|| < ||s(\omega)|| < \delta$, for all $(u, v) \in [0, 1] \times [0, 1]$, the integrand above is uniformly bounded and consequently by the Dominated Convergence Theorem, we can take the limit as $n \to \infty$ inside the integral. Thus we have

$$\begin{aligned} g(\gamma_i) - \hat{g}(\gamma_i) &= s(\omega)^T \left(\int_0^1 \int_0^1 v\ddot{g}\left(\hat{\gamma}_i(\omega) + uvs(\omega)\right) - \frac{1}{2}\ddot{g}\left(\gamma_i(\omega)\right) dudv \right) s(\omega) \\ &= o(||s(\omega)||) \to 0 \qquad n \to \infty \end{aligned}$$

This implies that

$$\hat{f}(Y_i, \gamma_i \mid \theta) - f(Y_i, \gamma_i \mid \theta) \xrightarrow{\text{a.s.}} 0.$$

□

Further discussion of the approximation error on this density are given Appendix A.3. Simulated results regarding the approximation error in the joint density are given below, in Section 4.4.2.

In this approximate form, we recognize that the second-order Taylor term for $g(\gamma_i)$ and the kernel for the density of $\gamma_i|\theta$, $p(\gamma_i \mid \theta)$, are both quadratic forms for γ_i. This allows us to apply the complete-the-square formula and rewrite the approximate joint density as

$$\begin{aligned}\hat{f}(Y_i, \gamma_i \mid \theta) &= \frac{h(Y_i, \theta)}{\sqrt{2\pi\sigma^2}} \exp\left(g(\hat{\gamma}_i) - \frac{1}{2}\left(-\ddot{g}(\hat{\gamma}_i)\,(\gamma_i - \hat{\gamma}_i)^2 + \frac{1}{\sigma^2}\,(\gamma_i - \psi_i)^2\right)\right) \\ &= \frac{h(Y_i, \theta)\exp\left(g(\hat{\gamma}_i)\right)}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\left(\frac{1}{\sigma^2} - \ddot{g}(\hat{\gamma}_i)\right)(\gamma_i - \gamma_i^*)^2 + u(Y_i, \theta)\,(\hat{\gamma}_i - \psi_i)^2\right)\right),\end{aligned}$$

where

$$\gamma_i^* = \frac{\psi_i - \sigma^2\ddot{g}(\hat{\gamma}_i)\hat{\gamma}_i}{1 - \sigma^2\ddot{g}(\hat{\gamma}_i)},$$

and

$$\begin{aligned}u(Y_i, \theta) &= (-\ddot{g}(\hat{\gamma}_i))\left(\frac{1}{\sigma^2} - \ddot{g}(\hat{\gamma}_i)\right)^{-1}\left(\frac{1}{\sigma^2}\right) \\ &= \frac{-\ddot{g}(\hat{\gamma}_i)}{\sigma^2\left(\frac{1}{\sigma^2} - \ddot{g}(\hat{\gamma}_i)\right)} \\ &= \frac{-\ddot{g}(\hat{\gamma}_i)}{1 - \sigma^2\ddot{g}(\hat{\gamma}_i)}\end{aligned}$$

We have now isolated γ_i into the term $\exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2} - \ddot{g}(\hat{\gamma}_i)\right)(\gamma_i - \gamma_i^*)^2\right)$. We recognize this as the kernel of a normal density with mean γ_i^* and variance

$$\frac{1}{\frac{1}{\sigma^2} - \ddot{g}(\hat{\gamma}_i)} = \frac{\sigma^2}{1 - \sigma^2\ddot{g}(\hat{\gamma}_i)}.$$

Then under our approximation, we say

$$\gamma_i \mid Y_i, \theta \sim \mathrm{N}\left(\gamma_i^*, \frac{\sigma^2}{1 - \sigma^2\ddot{g}(\hat{\gamma}_i)}\right).$$

We are now in a position to approximate our marginal likelihood.

Revisiting our approximation to the joint density, we have

$$\begin{aligned}
\hat{f}(Y_i, \gamma_i \mid \theta) &= \frac{h(Y_i, \theta) \exp\left(g(\hat{\gamma}_i)\right)}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{-\ddot{g}(\hat{\gamma}_i)}{1-\sigma^2\ddot{g}(\hat{\gamma}_i)}\right)(\hat{\gamma}_i - \psi_i)^2\right) \\
&\quad \times \exp\left(-\frac{1}{2}\left(\left(\frac{1}{\sigma^2} - \ddot{g}(\hat{\gamma}_i)\right)(\gamma_i - \gamma_i^*)^2\right)\right) \\
&= \frac{h(Y_i, \theta) \exp\left(g(\hat{\gamma}_i)\right)}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{-\ddot{g}(\hat{\gamma}_i)}{1-\sigma^2\ddot{g}(\hat{\gamma}_i)}\right)(\hat{\gamma}_i - \psi_i)^2\right) \\
&\quad \times \left(\frac{\sqrt{2\pi\left(\frac{\sigma^2}{1-\sigma^2\ddot{g}(\hat{\gamma}_i)}\right)}}{\sqrt{2\pi\left(\frac{\sigma^2}{1-\sigma^2\ddot{g}(\hat{\gamma}_i)}\right)}}\right) \exp\left(-\frac{1}{2}\left(\left(\frac{1}{\sigma^2} - \ddot{g}(\hat{\gamma}_i)\right)(\gamma_i - \gamma_i^*)^2\right)\right) \\
&= \frac{h(Y_i, \theta)\sqrt{2\pi\left(\frac{\sigma^2}{1-\sigma^2\ddot{g}(\hat{\gamma}_i)}\right)} \exp\left(g(\hat{\gamma}_i)\right)}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{-\ddot{g}(\hat{\gamma}_i)}{1-\sigma^2\ddot{g}(\hat{\gamma}_i)}\right)(\hat{\gamma}_i - \psi_i)^2\right) \\
&\quad \times \frac{1}{\sqrt{2\pi\frac{\sigma^2}{1-\sigma^2\ddot{g}(\hat{\gamma}_i)}}} \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2} - \ddot{g}(\hat{\gamma}_i)\right)(\gamma_i - \gamma_i^*)^2\right) \\
&= \frac{h(Y_i, \theta) \exp\left(g(\hat{\gamma}_i)\right)}{\sqrt{1-\sigma^2\ddot{g}(\hat{\gamma}_i)}} \exp\left(-\frac{1}{2}\left(\frac{-\ddot{g}(\hat{\gamma}_i)}{1-\sigma^2\ddot{g}(\hat{\gamma}_i)}\right)(\hat{\gamma}_i - \psi_i)^2\right) \\
&\quad \times \frac{1}{\sqrt{2\pi\frac{\sigma^2}{1-\sigma^2\ddot{g}(\hat{\gamma}_i)}}} \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2} - \ddot{g}(\hat{\gamma}_i)\right)(\gamma_i - \gamma_i^*)^2\right) \\
&= \hat{f}(Y_i \mid \theta) \times \hat{f}(\gamma_i \mid Y_i, \theta).
\end{aligned}$$

Here $\hat{f}(\gamma_i \mid Y_i, \theta)$ is the exact density for a $\mathrm{N}\left(\gamma_i^*, \frac{\sigma^2}{1-\sigma^2\ddot{g}(\hat{\gamma}_i)}\right)$ random variable. Then

$$\begin{aligned}
f(Y_i \mid \theta) &= \int f(Y_i, \gamma_i \mid \theta) d\gamma_i \\
&= \int \hat{f}(Y_i, \gamma_i \mid \theta) \exp\left(r(g, \gamma_i, \hat{\gamma}_i)\right) d\gamma_i \\
&= \int \hat{f}(Y_i \mid \theta)\hat{f}(\gamma_i \mid Y_i, \theta) \exp\left(r(g, \gamma_i, \hat{\gamma}_i)\right) d\gamma_i \\
&= \hat{f}(Y_i \mid \theta) \int \hat{f}(\gamma_i \mid Y_i, \theta) \exp\left(r(g, \gamma_i, \hat{\gamma}_i)\right) d\gamma_i \\
&\simeq \hat{f}(Y_i \mid \theta) \int \hat{f}(\gamma_i \mid Y_i, \theta) d\gamma_i \\
&= \hat{f}(Y_i \mid \theta).
\end{aligned}$$

This, finally, yields our approximate marginal density for REO GLMMs with a single random effect:

$$\hat{f}(Y_i \mid \theta) = \frac{h(Y_i, \theta) \exp\left(g(\hat{\gamma}_i)\right)}{\sqrt{1 - \sigma^2 \ddot{g}(\hat{\gamma}_i)}} \exp\left(-\frac{1}{2}\left[\frac{-\ddot{g}(\hat{\gamma}_i)}{1 - \sigma^2 \ddot{g}(\hat{\gamma}_i)} (\hat{\gamma}_i - \psi_i)^2\right]\right). \tag{4.10}$$

Earlier, we mentioned the approach of Cai and Dunson (2006). They approximate the marginal by taking a second-order Taylor series expansion of $L(\theta, \boldsymbol{\gamma} \mid \mathbf{Y}) \propto f(\mathbf{Y} \mid \boldsymbol{\gamma}, \theta)$ in γ around $\boldsymbol{\gamma} = 0_{qk}$. They then take the expectation of their Taylor expansion over the modeled random effects distribution. Their method results in an approximation of the form

$$\tilde{L}(\theta \mid \mathbf{Y}) = f(\mathbf{Y} \mid 0_{qk}, \theta) + \frac{1}{2} f(\mathbf{Y} \mid 0_{qk}, \theta) \operatorname{tr}\left(\mathbf{Z}^T V \mathbf{Z} \boldsymbol{\Sigma}\right),$$

where $\eta = \mathbf{X}\beta + \mathbf{Z}\boldsymbol{\gamma}$ and

$$V = \frac{\partial^2}{\partial \eta \partial \eta^T} \log f(\mathbf{Y} \mid 0_{qk}, \theta) + \operatorname{Diag}\left(\frac{\partial^2}{\partial \eta \partial \eta^T} \log f(\mathbf{Y} \mid 0_{qk}, \theta)\right).$$

The precise analytic statements of their approximation for the binomial and Poisson are mysterious and complex.

In contrast, we approximate the marginal for each cluster by taking a second-order Taylor expansion of $g(\gamma_i)$, the exponential part of $f(Y_i \mid \gamma_i, \theta)$. This expansion is taken around an offset GLM estimate for γ_i, treating it as if it were a parameter with β fixed (see Section 5.2.1.1). This is likely to be much closer to the true value of γ_i for a cluster than the Cai and Dunson pivot, the zero vector. We marginalize over $\boldsymbol{\gamma}$ by recognizing that, with normal random effects, the approximate conditional distribution of $\boldsymbol{\gamma}$ is multivariate normal when the Taylor remainder is ignored. The integral of this density is necessarily 1, and so it is easily removed, leaving us with an approximate marginal.

Cai and Dunson (2006) also assume their random effects follow a normal distribution, so they and we consider marginalizing within the same class of models. Our methods, in this chapter and the next, give substantially simpler computational forms. In Chapter 6, we discuss our plans to directly compare our approximation to theirs.

4.3.2 The Binomial and Poisson Approximations

In this subsection, we give precise specifications of key results from Section 4.3.1 as they pertain to the binomial and Poisson GLMMs, with some discussion of hypothetical scenarios in which these models can be used.

4.3.2.1 Binomial data

An REO binomial model is, in essence, just a model where each binomial observation with its collection of covariates is given an individual random effect. Since the sum of *iid* binomial random variables is a binomial random variable with the same success probability, our scenario of having multiple *iid* observations within each cluster is the same as having one observation on a larger binomial. Having only one binomial observation is not a hindrance to the use of this model, as long as m_{ij}, the number of observed Bernoulli trials for the i^{th} cluster is sufficiently large. In this case, by sufficiently large we mean that (1) we have enough data for the Taylor remainder term to be small, and (2) for all i, $\sum_{j=1}^{n} y_{ij}$ must not be equal to 0 or $\sum_{j=1}^{n} m_{ij}$, otherwise the approximation for GLM-type models will fail.

In our work on this method, we assume $m_{ij} = 1$ and n large instead of assuming m_{ij} large and $n = 1$. We do this because we find that this helps us think about the role the REO constraint plays here. The model must have multiple response observations where covariate combinations are constant within each cluster.

By way of example, consider a longitudinal data set in which ovulation status is recorded on k women over n menstrual cycles. If covariate data include only demographic information on the participants (e.g. race, educational attainment), the REO binomial model would be a reasonable approach to these data. If observation-related covariates were also desired (e.g. pollutant exposure levels during cycle), it would be inappropriate to use the marginalization approximation below.

We begin by defining a number of the key quantities in the approximation for binomial data.

$$\begin{aligned} y_{ij} &\sim \text{Bin}\left(m_{ij}, p_i\right) \\ p_i &= \frac{\exp\left(x_i\beta + \gamma_i\right)}{1 + \exp\left(x_i\beta + \gamma_i\right)} \\ h(Y_i, \theta) &= \prod_{j=1}^{n} \binom{m_{ij}}{y_{ij}} \\ \hat{\gamma}_i &= \log\left(\frac{\sum_{j=1}^{n} y_{ij}}{\sum_{j=1}^{n}\left(m_{ij} - y_{ij}\right)}\right) - x_i\beta \end{aligned}$$

We further define the total number of Bernoulli trials in cluster i, $m_{i\cdot} = \sum_{j=1}^{n} m_{ij}$, and the total number of observed events in cluster i, $y_{i\cdot} = \sum_{j=1}^{n} y_{ij}$. This allows us to simplify the definition of $\hat{\gamma}_i$ to

$$\hat{\gamma}_i = \log\left(\frac{y_{i\cdot}}{m_{i\cdot} - y_{i\cdot}}\right) - x_i\beta.$$

The functions $g(\hat{\gamma}_i)$ and $\ddot{g}(\hat{\gamma}_i)$, and their simplifications, require a bit more algebra. First, $g(\hat{\gamma}_i)$ is given by

$$\begin{aligned}
g(\hat{\gamma}_i) &= (x_i\beta + \hat{\gamma}_i)\sum_{j=1}^{n} y_{ij} - \log\left(1 + \exp\left(x_i\beta + \hat{\gamma}_i\right)\right)\sum_{j=1}^{n} m_{ij} \\
&= \log\left(\frac{y_{i\cdot}}{m_{i\cdot} - y_{i\cdot}}\right) y_{i\cdot} - \log\left(1 + \left(\frac{y_{i\cdot}}{m_{i\cdot} - y_{i\cdot}}\right)\right) m_{i\cdot} \\
&= \log\left(\frac{y_{i\cdot}}{m_{i\cdot} - y_{i\cdot}}\right) y_{i\cdot} - \log\left(\frac{m_{i\cdot}}{m_{i\cdot} - y_{i\cdot}}\right) m_{i\cdot} \\
&= \log\left(y_{i\cdot}\right) y_{i\cdot} - \log\left(m_{i\cdot}\right) m_{i\cdot} \\
&\qquad - \left(\log\left(m_{i\cdot} - y_{i\cdot}\right) y_{i\cdot} - \log\left(m_{i\cdot} - y_{i\cdot}\right) m_{i\cdot}\right) \\
&= y_{i\cdot}\log y_{i\cdot} + \left(m_{i\cdot} - y_{i\cdot}\right)\log\left(m_{i\cdot} - y_{i\cdot}\right) - m_{i\cdot}\log m_{i\cdot}.
\end{aligned}$$

Then $\ddot{g}(\hat{\gamma}_i)$ is given by

$$\begin{aligned}
\ddot{g}(\hat{\gamma}_i) &= -\left(\frac{\exp\left(x_i\beta + \hat{\gamma}_i\right)}{1 + \exp\left(x_i\beta + \hat{\gamma}_i\right)}\right)\left(\frac{1}{1 + \exp\left(x_i\beta + \hat{\gamma}_i\right)}\right)\sum_{j=1}^{n} m_{ij} \\
&= -\left(\frac{\frac{y_{i\cdot}}{m_{i\cdot} - y_{i\cdot}}}{1 + \frac{y_{i\cdot}}{m_{i\cdot} - y_{i\cdot}}}\right)\left(\frac{1}{1 + \frac{y_{i\cdot}}{m_{i\cdot} - y_{i\cdot}}}\right) m_{i\cdot} \\
&= -\left(\frac{\frac{y_{i\cdot}}{m_{i\cdot} - y_{i\cdot}}}{\frac{m_{i\cdot}}{m_{i\cdot} - y_{i\cdot}}}\right)\left(\frac{1}{\frac{m_{i\cdot}}{m_{i\cdot} - y_{i\cdot}}}\right) m_{i\cdot} \\
&= -\left(\frac{y_{i\cdot}}{m_{i\cdot}}\right)\left(\frac{m_{i\cdot} - y_{i\cdot}}{m_{i\cdot}}\right) m_{i\cdot} \\
&= -\frac{y_{i\cdot}\left(m_{i\cdot} - y_{i\cdot}\right)}{m_{i\cdot}}
\end{aligned}$$

Finally, plugging into Equation (4.8) gives us an approximate marginal density for the REO binomial GLMM.

$$
\begin{aligned}
&\hat{f}(\mathbf{Y} \mid \theta) \\
&= \prod_{i=1}^{k} \left[\frac{h(Y_i, \theta) \exp\left(g(\hat{\gamma}_i)\right)}{\sqrt{1 - \sigma^2 \ddot{g}(\hat{\gamma}_i)}} \exp\left(-\frac{1}{2} \left[\frac{-\ddot{g}(\hat{\gamma}_i)}{1 - \sigma^2 \ddot{g}(\hat{\gamma}_i)} \left(\hat{\gamma}_i - \psi_i\right)^2 \right] \right) \right] \\
&= \prod_{i=1}^{k} \left[\frac{\left(\prod_{j=1}^{n} \binom{m_{ij}}{y_{ij}} \right) \exp\left(y_{i\cdot} \log y_{i\cdot} + (m_{i\cdot} - y_{i\cdot}) \log\left(m_{i\cdot} - y_{i\cdot}\right) - m_{i\cdot} \log m_{i\cdot} \right)}{\sqrt{1 + \frac{\sigma^2 y_{i\cdot}(m_{i\cdot} - y_{i\cdot})}{m_{i\cdot}}}} \right] \\
&\qquad \times \exp\left(-\frac{1}{2} \sum_{i=1}^{k} \left[\frac{\frac{y_{i\cdot}(m_{i\cdot} - y_{i\cdot})}{m_{i\cdot}}}{1 + \frac{\sigma^2 y_{i\cdot}(m_{i\cdot} - y_{i\cdot})}{m_{i\cdot}}} \left(\log\left(\frac{y_{i\cdot}}{m_{i\cdot} - y_{i\cdot}} \right) - \left(x_i \beta + \psi_i\right) \right)^2 \right] \right) \\
&= \prod_{i=1}^{k} \left[\frac{\left(\prod_{j=1}^{n} \binom{m_{ij}}{y_{ij}} \right) \exp\left(y_{i\cdot} \log y_{i\cdot} + (m_{i\cdot} - y_{i\cdot}) \log\left(m_{i\cdot} - y_{i\cdot}\right) - m_{i\cdot} \log m_{i\cdot} \right)}{\sqrt{\frac{m_{i\cdot} + \sigma^2 y_{i\cdot}(m_{i\cdot} - y_{i\cdot})}{m_{i\cdot}}}} \right] \\
&\qquad \times \exp\left(-\frac{1}{2} \sum_{i=1}^{k} \left[\frac{y_{i\cdot}\left(m_{i\cdot} - y_{i\cdot}\right)}{m_{i\cdot} + \sigma^2 y_{i\cdot}\left(m_{i\cdot} - y_{i\cdot}\right)} \left(\log\left(\frac{y_{i\cdot}}{m_{i\cdot} - y_{i\cdot}} \right) - \left(x_i \beta + \psi_i\right) \right)^2 \right] \right)
\end{aligned}
$$

4.3.2.2 Poisson data

As with the binomial model above, the REO Poisson model is functionally equivalent to a model where each Poisson observation has its own set of covariates and an individual random effect. This results from the fact that the sum of *iid* Poisson random variables is itself a Poisson random variable; and for our purposes, no information is being lost by combining them in this scenario[3]. One must still have sufficient data for the method to work, however. Just as the binomial GLMM fails when either all trials are successes or all trials are failures for a particular individual i, the Poisson GLMM fails when no events are observed for a particular individual.

Once again, we find it easier to consider the data in a separated form: where counts are recorded independently over fixed intervals, rather than being totaled across one larger in-

[3]Fisher information loss is historically an area of significant concern with the Poisson distribution—especially to French statisticians. It is the French who coined the famous phrase, "Only a Fisher understands the behavior of the Poisson."

terval. We do this by assuming $n > 1$ responses come from a $\text{Pois}(\lambda_i)$ distribution, rather than assuming $n = 1$ response comes from a $\text{Pois}(m_{ij}\lambda_i)$ distribution. This framework makes it easier to understand the role of repeated exchangeable observations in the model.

Consider data on hospital-acquired infections (HAIs) taken at k hospitals over a period of n months. The framework below provides approximate marginalization for a model where HAIs are predicted on the basis of hospital-level characteristics only (e.g. nurse-to-patient ratio, socioeconomic status of surrounding community); but will not provide approximate marginalization when observation-related covariates are also desired (e.g. month of observation).

For the REO Poisson GLMM, we define the following:

$$\begin{aligned} y_{ij} &\sim \text{Pois}\left(m_{ij}\lambda_i\right) \\ \lambda_i &= \exp\left(x_i\beta + \gamma_i\right) \\ h(Y_i, \theta) &= \prod_{j=1}^{n} \frac{m_{ij}^{y_{ij}}}{y_{ij}!} \\ \hat{\gamma}_i &= \log\left(\frac{\sum_{j=1}^{n} y_{ij}}{\sum_{j=1}^{n} m_{ij}}\right) - x_i\beta \end{aligned}$$

We define the totals $y_{i\cdot}$ and $m_{i\cdot}$ as in the previous section. The total number of observed events in cluster i is $y_{i\cdot} = \sum_{j=1}^{n} y_{ij}$. The total number of intervals observed in cluster i is $m_{i\cdot} = \sum_{j=1}^{n} m_{ij}$. This allows us to simplify the definition of $\hat{\gamma}_i$ to

$$\hat{\gamma}_i = \log\left(\frac{y_{i\cdot}}{m_{i\cdot}}\right) - x_i\beta.$$

Then $g(\hat{\gamma}_i)$ is given by

$$\begin{aligned} g(\hat{\gamma}_i) &= (x_i\beta + \hat{\gamma}_i)\sum_{j=1}^{n} y_{ij} - \exp(x_i\beta + \hat{\gamma}_i)\sum_{j=1}^{n} m_{ij} \\ &= \log\left(\frac{y_{i\cdot}}{m_{i\cdot}}\right) y_{i\cdot} - \frac{y_{i\cdot}}{m_{i\cdot}} m_{i\cdot} \\ &= \left(\log\left(\frac{y_{i\cdot}}{m_{i\cdot}}\right) - 1\right) y_{i\cdot} \end{aligned}$$

And $\ddot{g}(\hat{\gamma}_i)$ is given by

$$\begin{aligned} \ddot{g}(\hat{\gamma}_i) &= -\exp(x_i\beta + \hat{\gamma}_i)\sum_{j=1}^{n} m_{ij} \\ &= -\frac{y_{i\cdot}}{m_{i\cdot}} m_{i\cdot} \\ &= -y_{i\cdot} \end{aligned}$$

This gives us an approximate marginal density for the REO Poisson GLMM of

$$\begin{aligned} \hat{f}(\mathbf{Y} \mid \theta) &= \prod_{i=1}^{k}\left[\frac{h(Y_i,\theta)\exp(g(\hat{\gamma}_i))}{\sqrt{1-\sigma^2\ddot{g}(\hat{\gamma}_i)}}\exp\left(-\frac{1}{2}\left[\frac{-\ddot{g}(\hat{\gamma}_i)}{1-\sigma^2\ddot{g}(\hat{\gamma}_i)}(\hat{\gamma}_i-\psi_i)^2\right]\right)\right] \\ &= \prod_{i=1}^{k}\left[\frac{\left(\prod_{j=1}^{n}\frac{m_{ij}^{y_{ij}}}{y_{ij}!}\right)\exp\left(\left(\log\left(\frac{y_{i\cdot}}{m_{i\cdot}}\right)-1\right)y_{i\cdot}\right)}{\sqrt{1+\sigma^2 y_{i\cdot}}}\right] \\ &\qquad \times \exp\left(-\frac{1}{2}\sum_{i=1}^{k}\left[\frac{y_{i\cdot}}{1+\sigma^2 y_{i\cdot}}\left(\log\left(\frac{y_{i\cdot}}{m_{i\cdot}}\right)-(x_i\beta+\psi_i)\right)^2\right]\right) \end{aligned}$$

4.4 Simulation Results

In this section, we describe small-sample simulation results for our marginal approximation. We begin by describing our simulation procedures in detail. We then report results on approximation error for the joint distribution and for the marginal DIC calculations. Finally, we present results for a simulated variable selection task using a minimum DIC model selection criterion based on DIC_b, DIC_j, and DIC_b.

4.4.1 Description of Simulation Procedures

We conducted two sets of simulations to test our approximate marginal method for calculating the DIC. The first was designed to assess how accurate our marginalization method was in small samples. Because our approximation is only expected to behave well asymptotically, we must investigate how it behaves without large n. The second set of simulations was designed to assess the performance of the various DIC computation methods in an automated stepwise regression procedure, following from our argument in Section 3.3.2.2 that it is important to understand how a selection criterion performs in an unsupervised setting.

In the first simulation set for assessing approximation accuracy, data were generated under a range of choices for k (number of clusters) and n (number of observations per cluster). For a given k, a set of γ's were sampled from the quantiles of a standard normal distribution for the Bernoulli, and from a $\mathrm{N}(0, 1/4)$ for the Poisson. Quantile sampling for the γ's followed the scheme

$$\gamma_i = \Phi^{-1}\left(\frac{2i-1}{2k}\right), \qquad i \in \{1, ..., k\}.$$

We also sampled $x_{(i)}$'s from the quantiles of a standard normal, in the same fashion. The order of the $x_{(i)}$'s was then randomized and each individual i paired with one value to give us x_i, the covariate value measured on that individual. These were used to assemble X_i

matrices for each individual, with $X_i = J_n \otimes [\,1 \quad x_i\,]$, a vector of 1's for a grand mean and a vector of shared covariate values for each observation on an individual. We then generated response data, y, using quantiles of the Bernoulli[4] and Poisson distributions. For Bernoulli simulations $y_{ij} \sim \text{Bern}\,(p_i)$, with $p_i = \text{expit}\,(\beta_1 + x_i\beta_2 + \gamma_i)$ and $(\beta_1, \beta_2) = (0, 1)$. For Poisson simulations $y_{ij} \sim \text{Pois}\,(\lambda_i)$, with $\lambda_i = \exp(\beta_1 + x_i\beta_2 + \gamma_i)$ and $(\beta_1, \beta_2) = (3, {}^1\!/_4)$. Within each individual i, response data were generated using a modified quantile method. For Bernoulli data,

$$Y_i = \left\{0, F^{-1}_{\text{Bern}(p_i)}\left(\frac{3}{2n}\right), F^{-1}_{\text{Bern}(p_i)}\left(\frac{5}{2n}\right), ..., F^{-1}_{\text{Bern}(p_i)}\left(\frac{2n-3}{2n}\right), 1\right\},$$

and for Poisson data,

$$Y_i = \left\{0, F^{-1}_{\text{Pois}(\lambda_i)}\left(\frac{3}{2n}\right), F^{-1}_{\text{Pois}(\lambda_i)}\left(\frac{5}{2n}\right), ..., F^{-1}_{\text{Pois}(\lambda_i)}\left(\frac{2n-3}{2n}\right), F^{-1}_{\text{Pois}(\lambda_i)}\left(\frac{2n-1}{2n}\right)\right\}.$$

For each choice of k and n, data were simulated five times with random seeds selected by the authors for repeatability. This was done because, although the quantile method was used for most simulations here, pairings of γ_i's and x_i's remain random. Results for these simulations are presented in Sections 4.4.2 and 4.4.3.

In the second simulation set for assessing automated model selection behavior, we chose distinct (k, n) combinations after seeing our results from the first simulation set. Choices of k and n were made based on what we thought reasonable real-world data sets might look like; computational feasibility for repeated MCMC sampling; and our beliefs about how large k and n would need to be in order to obtain reasonable inferences for the model parameters. Automated model selection simulations were performed for Bernoulli response data with $(k = 20, n = 15)$ and $(k = 30, n = 15)$; and for Poisson response with $(k = 15, n = 10)$ and $(k = 20, n = 10)$. For each simulation, we created a design matrix including five covariates and an intercept: call them $\{x_0 \equiv 1, x_1, x_2, x_3, x_4, x_5\}$. For the Binomial simulations, we chose $\beta =$

[4]All quantiles of the Bern (p) are either 0 or 1, with quantiles up to $1 - p$ all equal to 0 and quantiles greater than $1 - p$ all equal to 1. Sampling n quantiles from a Bernoulli thus has the effect of giving the "correct" number of 0's and 1's that would be expected in n independent trials.

$[\,1/2 \quad 1 \quad 1/2 \quad 0 \quad 0 \quad 0\,]^T$; and for the Poisson simulations, we chose $\beta = [\,2 \quad 1/2 \quad 1/4 \quad 0 \quad 0 \quad 0\,]^T$. For each individual the covariate data were generated from a multivariate normal distribution,

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} \sim \mathrm{N}_5 \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1/3 & 0 \\ 0 & 0 & 1/3 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \right).$$

Values of the random effect γ_i were generated from a $\mathrm{N}(0, 1/4)$ distribution in all cases. Response data were generated randomly according to Bernoulli and Poisson distributions. As before, data were simulated with a new deterministic random seed in each simulation, to ensure that we obtained new-but-repeatable datasets each time.

Once the data were generated, OpenBUGS was used to obtain posterior samples of θ and γ under the full model where all covariates were included, as well as a DIC_b score for that model. Based on these posterior samples, joint and marginal DICs were then calculated in R. We performed a backwards stepwise procedure using DIC_b as a selection criterion. Beginning with the full model, at each step DIC_b scores were calculated for the current model and for each model where one covariate was removed[5]. The model with the smallest DIC_b was selected as a new current model, the associated covariate removed, and this process was repeated. The stepwise procedure was stopped when either the DIC for all reduced models exceeded the DIC of the current model, or when we reached a model containing only the intercept term. We logged the order of removal, the final model selected, and posterior means and standard deviations on the β coefficients for the final model. Then we performed the same procedure, beginning again from the full model, using a numerical approximation of DIC_j as a selection criterion. Finally, we performed the same procedure with our approximation

[5]We only considered the removal of x_2 through x_6. The intercept analogue covariate, x_1, was forced into each model.

to DIC_m as a selection criterion. Results for the automated model selection simulation are presented in Section 4.4.4.

All Bernoulli simulations and the first set of Poisson simulations in the stepwise variable selection task were performed with three chains, each with a burn-in period of 1000 iterations and included 5000 iterations in the posterior sample. Chains were thinned by a factor of $\lceil n/10 \rceil$ (for Bernoulli) or n (for Poisson) to reduce autocorrelation. This was done because our DIC methods rely on performing calculations on each of the posterior iterates; and having fewer, more nearly independent iterates improves the computational efficiency of our calculation methods. With $B_t = 5000$ posterior iterates, to numerically approximate DIC_m we must perform 5000 computations, be they quadrature or approximate marginalization. Not thinning gives us more information about the posterior distribution, but then requires more computation to obtain our DIC_m approximations.

The second set of Poisson simulations for the variable selection task was performed in the same way but using only one chain to increase the speed with which results could be obtained. We examined the convergence of three chains on a subset of the simulations we performed (2-3 combinations of k and n for each of the Bernoulli and Poisson schemes). Our thinning factor and number of burn-in iterations were chosen to ensure good convergence properties, based on the simulations examined.

4.4.2 Pseudo-KL Results

In Table 4.1, we present results showing how close our approximation of the joint distribution comes to the true joint distribution, using a measure we call the pseudo-Kullback-Leibler (pKL) divergence. Recall from Equation (3.1) that the generic μ_0-directed KL divergence is defined as

$$KL_{(\mu_0:\mu_1)} = \mathrm{E}_{\mu_0}[f_0(y)] - \mathrm{E}_{\mu_0}[f_1(y)]\,.$$

In our case, we know the true joint posterior for $\boldsymbol{\gamma}$ and θ, $P(\boldsymbol{\gamma}, \theta \mid \mathbf{Y})$. We also have a random sample from this posterior, $\{(\gamma^{(1)}, \theta^{(1)}), ..., (\gamma^{(B_t)}, \theta^{(B_t)})\}$. Thus for arbitrary μ, say $P^*(\boldsymbol{\gamma}, \theta \mid \mathbf{Y})$, we can use

$$\frac{1}{B_t} \sum_{s=1}^{B_t} \left(\log p(\boldsymbol{\gamma}^{(s)}, \theta^{(s)} \mid \mathbf{Y}) - \log p^*(\boldsymbol{\gamma}^{(s)}, \theta^{(s)} \mid \mathbf{Y}) \right)$$

to approximate the corresponding KL divergence.

In our case, we wish to assess to what extent our approximation of the joint posterior, $\hat{P}(\boldsymbol{\gamma}, \theta \mid \mathbf{Y})$, diverges from the true joint posterior. Our approximation simply replaces the exponent of $f(\mathbf{Y} \mid \boldsymbol{\gamma}, \theta)$ with the second-order Taylor approximation, dropping the remainder. Performing this approximation without adjusting the constant of integration, however, means that $\hat{P}(\cdot)$ is not a proper probability density. As we discuss in Appendix A.3, we cannot be sure that these approximate joint densities are always integrable in small samples, though we would be very surprised to find that they weren't.

This is important because $\hat{P}(\cdot)$ not being a probability density function means the KL divergence is no longer guaranteed to be non-negative. Because of this, a simple average of the difference in log densities[6] does not necessarily indicate how close the approximation comes to the truth. Instead, we use a measure of our own devising, the pKL, defined computationally as follows:

$$pKL(p, \hat{p}) = \frac{1}{B_t} \sum_{s=1}^{B_t} \left| \log p(\boldsymbol{\gamma}^{(s)}, \theta^{(s)} \mid \mathbf{Y}) - \log \hat{p}(\boldsymbol{\gamma}^{(s)}, \theta^{(s)} \mid \mathbf{Y}) \right|.$$

It is difficult to know what constitutes a large pKL score. As explained in the preceding chapter, the Kullback-Leibler divergence is closely related to information criteria for model selection, however, and so rules of thumb about what constitutes a large difference in AIC or DIC are likely to be reasonable here as well. SBCV suggest that a difference in DIC scores

[6]We use the word 'density' loosely here. In nearly all respects we treat $\hat{P}(\cdot)$ as if it were a real density, and will refer to it as such. The distinction is primarily of interest in relation to the KL divergence alone, owing to the fact that since our approximation is so similar to the true joint density, it is easy to obtain negative values here because of the small error in constant of integration. In no other area of our work have we found it necessary to treat $\hat{p}(\cdot)$ as anything but a probability density.

Bernoulli			
k	n	pKL	$2pKL/\overline{D(\theta)}_j$
10	5	0.584	1.51%
	10	0.508	0.71%
	15	0.441	0.43%
	25	0.330	0.20%
	50	0.223	0.07%
20	5	1.124	1.71%
	10	1.329	0.93%
	15	1.078	0.53%
	25	0.671	0.21%
	50	0.465	0.08%
30	5	1.009	1.10%
	10	2.306	1.14%
	15	1.413	0.47%
	25	1.066	0.22%
	50	0.688	0.08%

Poisson			
k	n	pKL	$2pKL/\overline{D(\theta)}_j$
10	5	0.555	0.79%
	10	0.416	0.30%
	15	0.345	0.17%
	25	0.281	0.08%
20	5	0.768	0.54%
	10	0.697	0.25%
	15	0.586	0.14%
	25	0.467	0.07%

Table 4.1: Pseudo-Kullback-Leibler divergences between $p(\boldsymbol{\gamma}, \boldsymbol{\theta} \mid \mathbf{Y})$ and $\hat{p}(\boldsymbol{\gamma}, \theta \mid \mathbf{Y})$, for simulated Bernoulli and Poisson response data.

of less than 3 has questionable utility. What pKL gives us can be thought of as a quick-and-dirty estimate of how much error we may be introducing into our DIC scores through our method of approximation. Notably, pKL is smaller than 3 in all cases, usually smaller than 1.5, and generally decreases as more observations are taken on each cluster.

Another useful way to think about pKL is in relation to the size of $p(\boldsymbol{\gamma}, \theta \mid \mathbf{Y})$. Recall that the joint calculation of $\overline{D(\theta)}$ is given by

$$\overline{D(\theta)}_j = -\frac{2}{B_t} \sum_{s=1}^{B_t} \log p(\boldsymbol{\gamma}^{(s)}, \theta^{(s)} \mid \mathbf{Y}).$$

Then the fraction

$$\frac{2pKL(p,\hat{p})}{\overline{D(\theta)}_j} = \frac{\frac{2}{B_t}\sum_{s=1}^{B_t}\left|\log p(\boldsymbol{\gamma}^{(s)},\theta^{(s)} \mid \mathbf{Y}) - \log \hat{p}(\boldsymbol{\gamma}^{(s)},\theta^{(s)} \mid \mathbf{Y})\right|}{-\frac{2}{B_t}\sum_{s=1}^{B_t}\log f(\boldsymbol{\gamma}^{(s)},\theta^{(s)} \mid \mathbf{Y})}$$
$$= \frac{\sum_{s=1}^{B_t}\left|\log p(\boldsymbol{\gamma}^{(s)},\theta^{(s)} \mid \mathbf{Y}) - \log \hat{p}(\boldsymbol{\gamma}^{(s)},\theta^{(s)} \mid \mathbf{Y})\right|}{-\sum_{s=1}^{B_t}\log p(\boldsymbol{\gamma}^{(s)},\theta^{(s)} \mid \mathbf{Y})}$$

gives pKL as a proportion of the average joint deviance. Using this, we can see that the approximation error relative to the average deviance decreases quickly as n increases, and is never above 2% of the average deviance, even in the smallest datasets we simulate.

Results in this section and the next are based on only five simulations, but our simulation design discussed above ensures perfect simulation of all relevant quantities. The only places where randomness occurs in these simulations are (1) in the pairing of an X_i value with a true γ_i value, and (2) in the OpenBUGS posterior sampling of the parameters. Standard deviations across these five simulations, for all quantities reported in this section and the next, are very small.

4.4.3 Comparing the Approximation to Numerical Integration

Of particular interest to us is how well our method approximates a true marginal DIC value for REO GLMMs. Clearly, if such a marginal DIC were easy to obtain, our method would have little value. Our comparisons focus on our approximation method and Gaussian quadrature. Approximation through MCMC integration was attempted, but proved so computationally prohibitive that we do not report results from that method.

We determined that Gauss-Legendre quadrature performed better than Gauss-Hermite quadrature in our simulations (i.e. required calculation of fewer nodes before the integral approximation converged). As discussed in Section 4.2.2, Gauss-Legendre quadrature is designed for integrating functions on the range $(-1, 1)$ and uses Legendre polynomials to appropriately

weight the information from different nodes. Because of the integration range for the Gauss-Legendre approach, it was necessary for us to use a change-of-variables substitution for γ. Properly, we replaced each γ_i with ν_i where

$$\nu_i = 2\left(\frac{\exp(\gamma_i)}{1+\exp(\gamma_i)} - \frac{1}{2}\right), \qquad \gamma_i = \log\left(\frac{1+\nu_i}{1-\nu_i}\right).$$

Then our joint density for an individual i, in (Y_i, ν_i) as opposed to (Y_i, γ_i), is given by

$$f_\nu(Y_i, \nu_i \mid \theta) = f_\gamma\left(Y_i, \log\left(\frac{1+\nu_i}{1-\nu_i}\right) \mid \theta\right)\frac{2}{1-\nu_i^2}.$$

Tables 5.2 and 5.3 give results for p_D, $\overline{D(\theta)}$, and DIC for each of four different computation methods. These are the: the native computations provided by OpenBUGS[7], computations based on the joint likelihood for θ and γ, quadrature calculations for the marginal likelihood of θ, and calculations based on our approximation to the marginal likelihood. Table 5.2 provides simulation results for the Bernoulli GLMM described above, and Table 5.3 gives those for the Poisson GLMM.

Of primary interest in these tables are the columns representing marginal computations using quadrature and using our approximation. We see that results are inconsistent when n is very small, especially in the computation of p_D, for both the Bernoulli and the Poisson simulations. Computations of p_D under our approximation are stable as n changes, however. This suggests that the differences between quadrature and approximation methods in these cases may depend more on issues with the quadrature computations. Quadrature methods are guaranteed to converge to the true integral of a function as the number of nodes increases, for all polynomial functions; and since $g(\cdot)$ is continuous in γ, it is well approximated by an arbitrarily large polynomial function. Thus, the quadrature method works if enough nodes are used—but the behavior of the quadrature computations in these settings suggests one

[7]Specifically OpenBUGS version 3.2.3 rev 1012.

		p_D				$\overline{D(\theta)}$				DIC			
k	n	BUGS	Joint	Quad	Appx	BUGS	Joint	Quad	Appx	BUGS	Joint	Quad	Appx
	5	4.8	6.2	1.6	2.1	64.1	77.3	66.5	66.1	68.9	83.5	68.1	68.2
	10	6.4	7.3	1.5	2.5	128.0	142.0	132.7	132.9	134.4	149.3	134.2	135.3
10	15	8.2	11.3	2.5	2.5	181.5	205.5	192.7	192.1	189.7	216.8	195.3	194.5
	25	9.0	12.1	2.4	2.3	294.3	322.3	311.8	311.3	303.3	334.4	314.2	313.6
	50	9.5	12.3	2.4	2.3	583.8	612.0	607.4	607.1	593.2	624.3	609.8	609.5
	5	4.6	7.3	-3.5	2.2	133.5	130.8	129.8	133.2	138.2	138.0	126.3	135.5
	10	11.2	18.7	2.7	2.4	253.3	286.4	264.5	261.6	264.4	305.1	267.3	264.0
20	15	15.0	21.9	2.8	2.3	361.9	410.1	383.8	381.1	376.9	431.9	386.6	383.4
	25	16.6	22.4	2.8	2.6	600.0	649.2	629.8	628.2	616.6	671.6	632.6	630.8
	50	18.4	22.7	2.8	2.7	1173.8	1228.9	1219.4	1218.4	1192.2	1251.5	1222.3	1221.1
	5	4.4	15.8	-5.2	2.3	199.7	183.1	193.7	199.2	204.1	198.9	188.5	201.5
	10	12.2	23.0	1.3	2.6	377.2	406.0	386.7	383.1	389.4	429.0	388.0	385.7
30	15	19.3	31.3	3.0	2.6	546.5	605.6	571.9	568.5	565.8	636.9	574.8	571.1
	25	24.2	32.6	2.9	2.6	874.2	948.4	918.2	915.4	898.4	981.0	921.1	918.0
	50	27.1	32.8	2.9	2.7	1724.9	1804.4	1789.7	1788.0	1752.0	1837.2	1792.6	1790.7

Table 4.2: Comparison of p_D's and DIC's for various calculation methods on a Bernoulli GLMM with random effects variance $\sigma^2 = 1$. Quadrature is based on a 50-node calculation.

		p_D				$\overline{D(\theta)}$				DIC			
k	n	BUGS	Joint	Quad	Appx	BUGS	Joint	Quad	Appx	BUGS	Joint	Quad	Appx
	5	6.4	8.8	409.6	2.8	133.4	141.0	547.4	140.4	139.8	149.8	957.0	143.2
	10	7.7	11.6	2.8	2.7	268.4	280.8	280.8	280.5	276.1	292.4	283.6	283.3
10	15	8.4	12.1	2.6	2.6	401.2	416.2	418.0	417.8	409.6	428.3	420.7	420.4
	25	8.9	12.3	2.6	2.6	668.1	683.2	689.1	688.9	677.0	695.5	691.7	691.5
	5	9.3	18.5	32.8	3.1	277.3	285.6	317.7	287.8	286.5	304.1	350.5	290.9
	10	13.1	21.7	2.9	2.9	528.1	548.7	549.6	548.9	541.2	570.3	552.5	551.8
20	15	14.9	22.1	2.9	2.9	802.2	824.9	830.5	830.0	817.1	847.1	833.4	832.9
	25	16.7	22.4	2.8	2.8	1326.3	1351.4	1363.8	1363.4	1343.0	1373.8	1366.6	1366.2

Table 4.3: Comparison of p_D's and DIC's for various calculation methods on a Poisson GLMM with random effects variance $\sigma^2 = {}^1/_4$. Quadrature is based on a 200-node calculation.

of the problems with using quadrature-based marginalization. The number of quadrature nodes necessary to give a good approximation to the marginal depends on the dataset and the log-likelihood function.

Because of the inconsistency in marginalized results for very small n, we are unable to say whether our approximation is accurate in those cases. We are also not able to say whether our approximation is inaccurate, because these comparisons are between our approximation and approximation by quadrature. No gold standard is available; we only know that the two approximations do not agree for $n = 5$. Regardless of where the inaccuracies arise from, we do not recommend that our approximation be used when five or fewer observations are available per cluster.

For $n \geq 15$, we see very good agreement between our method and quadrature for approximating all of p_{Dm}, $\overline{D(\theta)}_m$, and DIC_m. In the Bernoulli simulations, differences in $\overline{D(\theta)}_m$ and DIC_m are rarely greater than 3—and as Figure 4.1 shows below, those differences become very small relative to the overall size of the DIC_m statistic. In the Poisson simulations, very good agreement is also achieved at $n = 10$, and DIC_m differences are rarely more than 0.5.

Also of interest here is how the BUGS and joint DIC computations compare to the marginal computations. We are unsurprised to see that both BUGS and joint computation methods give a higher value for p_D than our method, based on our demonstration in Section 3.2.2. Of perhaps more interest is the behavior we see in $\overline{D(\theta)}$. The model fit value as assessed by this quantity varies as a function of k and n, as well as varying by computation method. It is not clear how best to compare $\overline{D(\theta)}$ across computation methods, meaning that while we can in some sense understand the meaning of the p_D differences, we cannot necessarily understand the $\overline{D(\theta)}$ differences except in the relative sense of how the various criteria perform at a model selection task. The important feature of these DIC computation methods is how well they function for model selection, and we cannot understand the differences between them except in the context of such a task. We return to this question in Section 4.4.4.

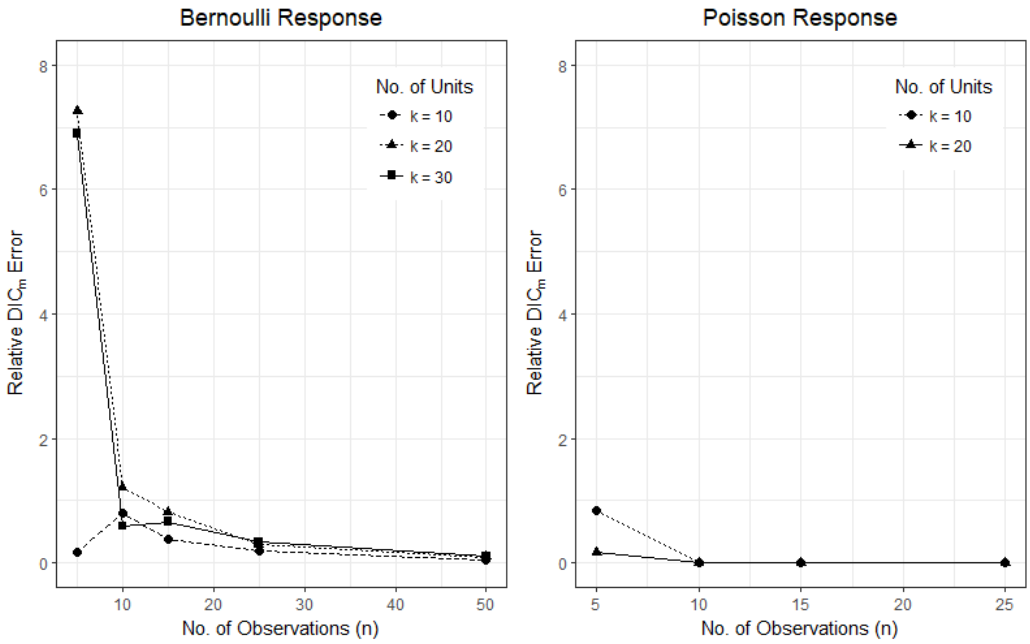

Figure 4.1: Absolute percentage difference in DIC_m under approximation and quadrature, relative to quadrature DIC_m.

Figure 4.1 shows the percentage difference in DIC_m computations, of the form

$$\left| \frac{DIC_{m,A} - DIC_{m,Q}}{DIC_{m,Q}} \right| * 100\%,$$

where $DIC_{m,A}$ is DIC_m computed using our approximation and $DIC_{m,Q}$ is DIC_m computed using Gaussian quadrature. Note that the percentage difference in the computations generally decreases as n increases, for all choices of k. From this figure, we can see that for $n > 10$ with Bernoulli response data, there is less than 0.75% difference in our computations. The Poisson response data shows even less relative difference in DIC_m computations, with no appreciable error as early as $n = 10$. As previously discussed, inconsistencies in the quadrature computations at $n = 5$ make it difficult to interpret the differences shown at that level.

[8]The simulations giving these values were run on home laptop computers. We have no information about the provenance of this number, but we assume it must have occurred because the laptop running these simulations suspended the simulation process for around 4 hours and 37 minutes.

Bernoulli

k	n	BUGS	Joint	Q-50	Appx
10	5	15	1	18	1
	10	26	1	19	1
	15	67	1	19	1
	25	159	1	18	1
	50	523	1	19	2
20	5	53	1	37	2
	10	99	1	37	2
	15	279	1	37	2
	25	687	1	36	2
	50	2336	2	37	3
30	5	123	1	54	2
	10	245	1	54	2
	15	704	2	54	3
	25	1734	2	54	4
	50	6863	3	55	6

Poisson

k	n	BUGS	Joint	Q-200	Appx
10	5	67	1	226	2
	10	279	1	227	2
	15	655	1	227	2
	25	2056	1	232	2
20	5	1121	1	445	3
	10	5996	1	451	3
	15	17287	1	17073[8]	3
	25	77155	1	455	4

Table 4.4: Runtimes (in seconds) for DIC approximations. Since joint, quadrature, and approximation calculations are based on existing MCMC samples, each of these columns reports the additional time necessary for these calculations after the BUGS runtime has already elapsed.

Table 4.4 reports the runtimes of the various DIC calculation methods. Note that the number of quadrature nodes used for the Bernoulli simulations were $B_n = 50$, and for the Poisson we used $B_n = 200$. These were chosen based on our observations of how many nodes were required to achieve convergence to the integral in a subset of cases considered. We see that the joint and approximate marginal calculations execute very quickly, requiring only one algebraic calculation for each posterior iterate $\theta^{(s)}$. The quadrature calculations take appreciably longer, requiring B_n calculations per posterior iterate. BUGS runtimes are the longest for each of the scenarios presented, but these runtimes represent a fixed cost. The joint, quadrature, and approximate marginal calculations all require MCMC output, and so runtimes listed for them are additional runtimes necessary to get these values once the MCMC has been completed.

Because of the above results, as well as the necessity to transform $g(\cdot)$ for use with quadrature, and the longer runtime necessary for quadrature methods, we feel that our method provides a preferable alternative for marginalization in the REO setting. Implementing it is not appreciably more difficult, $p_{Dm,A}$ and $DIC_{m,A}$ are close to $p_{Dm,Q}$ and $DIC_{m,Q}$ respectively, and our approximation method takes less time for a computer to execute.

4.4.4 Simulated Stepwise Variable Selection

Our second set of simulations is intended to compare the behavior of an automated model selection procedure based on a minimum-DIC criterion using DIC_b, DIC_j, and DIC_m as defined in Section 3.3.1. Table 4.5 presents results for backwards stepwise procedures using 100 simulations of Bernoulli repeated-observations data with $k = 20$ and $n = 15$; 100 Bernoulli simulations with $k = 30$ and $n = 15$; 5 Poisson simulations with $k = 15$ and $n = 10$; and 8 Poisson simulations with $k = 20$ and $n = 10$.

We consider a number of metrics we think users may find useful:

Pr(Correct) The proportion of times the stepwise procedure includes precisely the variables in the correct (i.e. generating) model,

Pr(No O-Errors) The proportion of times the procedure includes each variable in the generating model, regardless of whether other variables are included,

O-Errors The average number of omission errors—how many variables in the generating model are not included in the final model,

I-Errors The average number of inclusion errors—how many variables not in the generating model are included in the final model.

GMR What we term the good model rate (GMR) for this problem—the proportion of times the final model includes less than two errors of either type.

From this table, we observe that DIC_b and DIC_m pick the correct model more often than DIC_j for the Bernoulli simulations, with DIC_m correct slightly more often than DIC_b as well. In the Poisson simulations, DIC_j and DIC_m are about equally successful, and both are better than DIC_b—though we must note that computation time for these models was high, and our results are based on only a small number of simulated stepwise procedures. Each DIC criterion is likely to lead to a model that includes all the relevant covariates. The average number of relevant covariates missed is low, regardless of criterion. We find, however, that the average number of spurious covariates included in the selected model tends to be lowest across all simulations for DIC_m. Our criterion, DIC_m, picks the correct model as or more often than either DIC_b or DIC_j, shows little drop-off in its tendency to pick relevant covariates, and shows appreciably less tendency to pick spurious covariates.

In assessing overall model quality, we find that the GMR tends to be highest for models chosen using our method. That is, in our simulations where the response-generating covariates

Model Type	k, n	Simulations	Method	Pr(Correct)	Pr(No O-Errors)	O-Errors	I-Errors	GMR
Bernoulli	20, 15	100[9]	BUGS	0.41	0.85	0.15	0.78	0.78
			Joint	0.11	0.87	0.14	1.61	0.41
			Marginal	0.46	0.82	0.19	0.63	0.83
	30, 15	100^6	BUGS	0.56	0.98	0.02	0.64	0.81
			Joint	0.14	0.95	0.05	1.64	0.40
			Marginal	0.59	0.96	0.04	0.50	0.87
Poisson	15, 10	5	BUGS	0.00	0.80	0.20	1.20	0.60
			Joint	0.20	0.80	0.20	1.00	0.60
			Marginal	0.20	0.80	0.20	1.00	0.60
	20, 10	8	BUGS	0.00	0.50	0.63	2.00	0.13
			Joint	0.63	0.63	0.38	0.38	0.75
			Marginal	0.63	0.63	0.38	0.38	0.75

Table 4.5: Comparison of backwards stepwise selection output for BUGS, joint, and marginal approximation methods of DIC.

[9] These simulations were performed on Windows machines using OpenBUGS 3.2.3. All other simulations were performed on a Linux server using OpenBUGS 3.2.3.

are known, the DIC_m criterion is more likely to pick the correct set of covariates than either competitor criterion, and it is more likely to pick a set of covariates near to the correct set than either competitor criterion.

4.5 Discussion

Based on the above results, we have argued that our method for marginal approximation in a GLMM with repeated measures data may be more likely to pick the correct model in an automated backwards stepwise model selection procedure. We do not think this should surprise our readers. As we discussed in the preceding chapter, when considering mixed models, the DIC makes the clearest sense when looking at a marginalized model. Standard BUGS DICs and joint DICs incorporate information on nuisance variables and artificially inflate the p_D component of DIC. Even in small samples, true distributions are well approximated by our method—both for the joint posterior of θ and γ (Table 4.1), and for the marginal distribution of y given θ (Tables 4.2 and 4.3).

Still, this method is somewhat limited because it requires us to have repeated exchangeable observations within each cluster and does not allow us to consider observation-level covariates. We address these issues in the next chapter, by introducing new tools that broaden our method to be used with all GLMMs.

Chapter 5

Marginalization for DIC – Part III

In Chapter 3 we introduced the mixed model, discussed three numerical approximation methods for the DIC (DIC_j, DIC_b, and DIC_m), and presented methods for marginalizing linear mixed models in the Bayesian setting. In chapter 4, we developed a novel method for approximate marginalization in repeated exchangeable observations (REO) GLMMs—GLMMs where only cluster-level covariates are considered. We explained the asymptotic behavior of our approximation and used simulations to examine its behavior in small samples. We found that our approximation is fast; gives an approximate joint density that is close to the true joint density; and results in numerical p_{Dm} and DIC_m approximations that are close to what we'd expect from other, more computationally intensive marginalization methods. We also found that, in an automated stepwise variable selection procedure, a criterion based on our DIC_m approximation resulted in preferable behavior relative to criteria based on numerical approximations to DIC_j and DIC_b.

Chapter 5 presents a modification of the Chapter 4 method that can be used to approximate DIC_m for the general class of GLMMs, not just REO GLMMs. We begin with an explanation of why our method from Chapter 4 is insufficient to deal with the general case. Development

of the general approximation method follows, along with a discussion of approximation error and specific formulas for the Binomial (logistic) and Poisson GLMMs. Our small-sample simulations focus on the case where a subject-level random effect is modeled, since the class of GLMMs with such effects is sufficiently broad to be interesting while still computationally tractable.

5.1 New Considerations

In Section 4.3 we presented a limited marginalization approach based on an approximation method for use with REO GLMMs. In this section, we discuss the differences between REO and non-REO GLMMs and explain why our limited approach from Chapter 4 is insufficient to approximate the joint and marginal densities in non-REO GLMMs.

5.1.1 REO vs. Non-REO GLMMs

Our developments in the last chapter focused on repeated exchangeable observations (REO) GLMMs. These are GLMMs for which covariate data is constant within clusters. An example of cluster-level covariate data are long-term demographic information on individuals (e.g. educational attainment, race, marital status), with those individuals acting as our clusters. These cluster-level covariates may change over time, but they are often treated as constant because it is assumed that they won't change over the course of the study. Because REO GLMMs include only cluster-level covariates, response values within clusters (e.g. $y_{i1}, ..., y_{in}$) can be permuted without changing any aspect of the resulting inference. This is the origin of the "repeated exchangeable observations" nomenclature.

Methodological work on the special case of REO GLMMs has a long history in statistics. Breslow and Clayton (1993) discuss the development of frequentist estimation methods in

this context, mentioning especially the work of Hinde (1982) on the Poisson REO GLMM and Crouch and Spiegelman (1990) on the binomial REO GLMM. As we discussed in Section 4.3.2, the Poisson and binomial REO GLMMs are equivalent to allowing each Poisson or binomial observation to have its own random effect. In effect, because sums of independent Poissons and independent binomials with equal probabilities are also Poisson and binomial, respectively, we can consider every cluster to have size one. These models allow us to consider a wide range of possible data, but they do not allow us the ability to consider more complex structure in the data.

Non-REO GLMMs are those GLMMs for which observations on a cluster cannot be considered repeated exchangeable observations. These models include covariates whose values vary within clusters. Since GLMMs are frequently used for longitudinal data analysis where observations within a cluster are taken at different times, this type of covariate is often called a time-varying covariate. An example of time-varying covariate data is short-term health information (e.g. heart rate, blood pressure, blood oxygenation). We expect these covariates to fluctuate over time, and some response measurements may be influenced by those fluctuations. If response measurements are influenced by these covariates, then the repeated observations within a cluster cannot be considered exchangeable.

With non-REO GLMMs, we move beyond equivalence to size-one clusters. We can now consider clusters where observations share some inherent similarity but are not composed of fully exchangeable observations, as in the case of time-varying covariates or multi-level clustering.

5.1.2 Why the Limited Approach Fails

Recall that our approximate marginalization method uses a Taylor expansion of the function $g(\gamma_i)$ around some value $\hat{\gamma}_i$. A second-order Taylor polynomial gives us a quadratic function

for γ_i, which can be combined with $P(\gamma_i \mid \theta)$ using the complete-the-square formula. Our objective is to isolate γ_i in the kernel of a normal density so that it can be easily separated from the rest of the approximate joint density.

Isolating γ_i in this way depends on having γ_i appear only in the quadratic term and in the remainder. Under Taylor's theorem, however,

$$g(\gamma_i) = g(\hat{\gamma}_i) + \dot{g}(\hat{\gamma}_i)(\gamma_i - \hat{\gamma}_i) + \frac{1}{2}\ddot{g}(\hat{\gamma}_i)(\gamma_i - \hat{\gamma}_i)^2 + r(g, \gamma, \hat{\gamma}).$$

We see that γ_i also appears in the first-order linear term. In the REO GLMM setting, we can easily choose $\hat{\gamma}_i$ satisfying $\dot{g}(\hat{\gamma}_i) = 0$, eliminating this term. Elimination of the linear term is more difficult in the non-REO GLMM setting, as we will now explain.

To understand the difficulty, it is instructive to look at what happens when we attempt to set $\dot{g}(\hat{\gamma}) = 0$ in the binomial-logistic GLMM. We continue to use our earlier notation, but we now let X_{ij} refer to a row vector from the fixed-effect design matrix associated with response observation j on unit i. Analogously, Z_{ij} will refer to a row vector from the cluster i submatrix of the random-effect design matrix[1].

Under the logistic model, the function $g(\gamma_i)$ is given by

$$g(\gamma_i) = \sum_{j=1}^{n} \left(y_{ij} \left(X_{ij}\beta + Z_{ij}\gamma_i \right) - m_{ij} \log \left(1 + \exp \left(X_{ij}\beta + Z_{ij}\gamma_i \right) \right) \right).$$

Then taking a derivative with respect to γ_i gives

$$\dot{g}(\gamma_i) = \sum_{j=1}^{n} Z_{ij} \left(y_{ij} - m_{ij} \frac{\exp \left(X_{ij}\beta + Z_{ij}\gamma_i \right)}{1 + \exp \left(X_{ij}\beta + Z_{ij}\gamma_i \right)} \right).$$

[1]We hope that this explanation is sufficient for the reader to understand the structure of Z_{ij}—but if it is not, we refer the reader to Section 5.2.1 where our notation is defined in greater detail.

Setting $\dot{g}(\hat{\gamma}_i) = 0$ in the context of the previous chapter's scenario allows for a substantial simplification. Assume constant covariate values within clusters, namely $X_{i*} = X_{i1} = ... = X_{in}$ and $Z_{i*} = Z_{i1} = ... = Z_{in}$. Then setting the above expression to zero and solving, we obtain

$$\text{expit}\left(X_{i*}\beta + Z_{i*}\gamma_i\right) m_{i\cdot} = y_{i\cdot}.$$

Taking the logit function of both sides gives

$$X_{i*}\beta + Z_{i*}\gamma_i = \text{logit}\left(\frac{\frac{y_{i\cdot}}{m_{i\cdot}}}{1 - \frac{y_{i\cdot}}{m_{i\cdot}}}\right),$$

which can be rewritten as

$$Z_{i*}\gamma_i = \text{logit}\left(\frac{y_{i\cdot}}{m_{i\cdot} - y_{i\cdot}}\right) - X_{i*}\beta.$$

This leads to our development shown in the previous chapter.

If we do not have this simplification, solving for $\hat{\gamma}_i$ is more involved. Instead of the reduced form above, we need to solve the following equation for γ_i:

$$\sum_{j=1}^{n} m_{ij} \frac{\exp\left(X_{ij}\beta + Z_{ij}\gamma_i\right)}{1 + \exp\left(X_{ij}\beta + Z_{ij}\gamma_i\right)} = \sum_{j=1}^{n} y_{ij}.$$

This expression involves the sum of many expit functions, each containing γ_i. The above approach fails to result in a simple expression. We need to find some other way to obtain $\hat{\gamma}_i$ satisfying $\dot{g}(\hat{\gamma}_i) = 0$.

5.2 A General Marginalization Approach

This section presents our general method for approximate marginalization of non-REO GLMMs, including a discussion of root-finding. We also discuss the amount of approx-

imation error created through the non-REO GLMM approximation method. Finally, we present formulas for non-REO GLMM approximation for binomial and Poisson response data.

5.2.1 Approximate Marginalization through Root-Finding and Taylor Expansion

As discussed in the preceding section, in order to choose $\hat{\gamma}_i$ satisfying $\dot{g}(\hat{\gamma}_i) = 0$ in the non-REO setting, it is necessary to use a numerical root-finding method. We choose the Newton-Rhapson algorithm because of its connection to the iteratively reweighted least squares method of model fitting for GLMs.

5.2.1.1 The Newton-Rhapson Algorithm

The Newton-Rhapson (NR) algorithm, also known as Newton's method, is a numerical method for finding the root of an equation. For our method, we will be concerned with finding $\hat{\gamma}$ satisfying $\dot{g}(\hat{\gamma}) = 0$, the conditional maximizer of the $g(\gamma \mid y, \theta)$ function. For the exponential family likelihoods we consider, $\ddot{g}(\cdot)$ will be strictly negative definite, ensuring that $\hat{\gamma}$ is a maximizer.

As explained in R. R. Christensen (1997), Newton's method proceeds by taking successive approximations to the root. Let $f(\cdot)$ be a scalar function for which the root x is desired—that is, we seek a solution to $f(x) = 0$. An initial value is chosen, $x^{(0)}$. Then, given $x^{(i)}$, the algorithm calculates $x^{(i+1)}$ through the formula

$$x^{(i+1)} = x^{(i)} - \left[\dot{f}\left(x^{(i)}\right)\right]^{-1} f\left(x^{(i)}\right). \tag{5.1}$$

This formula is obtained by observing that if $x^{(i)}$ and $x^{(i+1)}$ are close to each other, with $x^{(i+1)} - x^{(i)} = \delta_i$, then $f\left(x^{(i+1)}\right)$ will be approximately equal to $f\left(x^{(i)}\right)$ plus δ_i times the slope (gradient) of $f(\cdot)$ evaluated at $x^{(i+1)}$. That is,

$$f\left(x^{(i+1)}\right) \doteq f\left(x^{(i)}\right) + \dot{f}\left(x^{(i)}\right)\delta_i. \tag{5.2}$$

Since we are seeking the root of $f(\cdot)$, we set the left hand side of Equation 5.2 equal to 0 and solve for δ_i. Then $\delta_i = -\left[\frac{\partial}{\partial x} f\left(x^{(i)}\right)\right]^{-1} f\left(x^{(i)}\right)$ is the iterative adjustment we apply to $x^{(i)}$ in Equation 5.1 above.

Applying this to the problem of root-finding for $\dot{g}(\cdot)$, we add a small step-size adjustment factor as described by Wolfe (1969). This method, sometimes called relaxed Newton's method, is common in vector-valued applications as a way of protecting against "overshooting". The iterative equation we use for root-finding is then

$$\gamma^{(i+1)} = \gamma^{(i)} - \zeta\left[\ddot{g}\left(\gamma^{(i)}\right)\right]^{-1}\dot{g}\left(\gamma^{(i)}\right), \tag{5.3}$$

where $\zeta \in (0, 1]$ is our step-size adjustment factor.

Note that in our setting, this maximization problem can be restated in the form of a GLM model with an offset. In the next section, we explain this restatement of the problem and discuss why it is useful to our approximate marginalization method.

5.2.1.2 Newton-Rhapson through iteratively reweighted least squares

R. R. Christensen (1997) points out that fitting a GLM model with iteratively reweighted least squares (IRLS) is simply an application of the Newton-Rhapson algorithm to maximize the log-likelihood function for the model parameters. In this section, we explain how this

applies to our maximization problem. We begin by repeating our notation from previous chapters for clarity.

Let $\mathbf{Y} = \{Y_i\} = \{y_{ij}\}$ be a $kn \times 1$ vector of response data on clusters $i \in \{1, ..., k\}$, with $j \in \{1, ..., n\}$ observations per cluster. We use a balanced design with common n for all clusters to simplify some of the following linear algebra, but the results we obtain do not require this balance.

Let β be a $p \times 1$ vector of regression parameters. Let $\mathbf{X}$ be the $kn \times p$ design matrix for the regression parameters, $\mathbf{X}_i$ be the $n \times p$ block of the $\mathbf{X}$ matrix corresponding to cluster i, and X_{ij} be the $1 \times p$ row vector corresponding to the j^{th} observation on cluster i.

Let $\boldsymbol{\gamma} = \begin{bmatrix} \gamma_1^T & ... & \gamma_k^T \end{bmatrix}^T$ be a $kq \times 1$ vector of random effects, with γ_i the $q \times 1$ vector of random effects corresponding to cluster i. Let $\mathbf{Z}$ be the $kn \times kq$ block diagonal design matrix for the random effects. Let $\mathbf{Z}_i$ be the $n \times q$ submatrix of $\mathbf{Z}$ corresponding to its i^{th} diagonal block, and Z_{ij} be the $1 \times q$ row vector corresponding to the j^{th} row of the $\mathbf{Z}_i$ matrix.

Let $\boldsymbol{\psi} = \begin{bmatrix} \psi_1^T & ... & \psi_k^T \end{bmatrix}^T$ be the mean of the random effects vector $\boldsymbol{\gamma}$, and let $\boldsymbol{\Sigma}$ be block diagonal Σ_i, $i \in \{1, ..., k\}$ be the covariance matrix of the random effects. We assume that $\boldsymbol{\gamma} \sim \mathrm{N}_{kq}(\boldsymbol{\psi}, \boldsymbol{\Sigma})$, or equivalently $\gamma_i \overset{\text{indep}}{\sim} \mathrm{N}_q(\psi_i, \Sigma_i)$. We use θ to refer to the collection of parameters $\{\beta, \boldsymbol{\psi}, \boldsymbol{\Sigma}\}$.

We assume that the elements of $\mathbf{Y}$ follow an exponential family distribution with pdf

$$f(Y_i|\gamma_i, \theta) = h(Y_i) \exp\left(g^*(Y_i \mid \gamma_i, \theta)\right).$$

Although $g^*(Y_i \mid \gamma_i, \theta)$ is a function of Y_i, we treat it as a function of γ_i. We write

$$g(\gamma_i \mid Y_i, \theta) \propto g^*(Y_i \mid \gamma_i, \theta).$$

We further simplify our notation by abbreviating $g(\gamma_i) \equiv g(\gamma_i \mid Y_i, \theta)$. We define $g(\boldsymbol{\gamma}) \equiv g(\boldsymbol{\gamma} \mid \mathbf{Y}, \theta)$ analogously relative to the full density $f(\mathbf{Y} \mid \boldsymbol{\gamma}, \theta)$.

Observe that within a cluster i, finding $\hat{\gamma}_i$ such that $\dot{g}(\hat{\gamma}_i) = 0$ is equivalent to finding the MLE for a GLM model for the data Y_i where $\mathbf{X}_i\beta$ is a known offset to the linear term and $\mathbf{Z}_i\gamma_i$ act as the design matrix and parameter vector for a fixed effects model. It does not make sense to talk about an "MLE" for our random effects γ_i, but the $\hat{\gamma}_i$ solution that maximizes $g(\cdot)$ is precisely the same as what we would obtain as an MLE in this offset fixed effects model.

As an example, consider a binomial GLM for each Y_i with covariates $\mathbf{X}_i$ and $\mathbf{Z}_i$, and known β. Then

$$y_{ij} \overset{\text{iid}}{\sim} \text{Bin}\left(m_{ij}, p_{ij}\right) \qquad\qquad \log\left(\frac{p_{ij}}{1-p_{ij}}\right) - X_{ij}\beta = Z_{ij}\gamma_i$$

We have simply changed the link function for logistic model to accommodate a known linear offset $X_{ij}\beta$ for each observation, but we are still in the GLM setting. We are seeking to maximize the log-likelihood $\ell(\gamma_i \mid Y_i, \beta)$, where

$$L(\gamma_i \mid Y_i, \beta) = \prod_{j=1}^{n} \binom{m_{ij}}{y_{ij}} \text{expit}\left(X_{ij}\beta + Z_{ij}\gamma_i\right)^{y_{ij}} + \left(1 - \text{expit}\left(X_{ij}\beta + Z_{ij}\gamma_i\right)\right)^{m_{ij}-y_{ij}}.$$

NR is a natural choice for finding the $\hat{\gamma}_i$'s we wish to use. Moreover, because of the IRLS fitting method used for GLMs and the equivalence between $\hat{\gamma}_i$ and the MLE for the offset GLM we have discussed, NR estimates can be obtained from any conventional software that computes MLEs for generalized linear models. Essentially, we are finding $\hat{\gamma}_i$ by using standard frequentist computational methods.

Other root-finding methods could be used instead of Newton-Rhapson. We prefer the NR algorithm, however, because IRLS methods are already well known in the GLM setting and the functionality to compute MLEs for a GLM model with an offset is widely available in current statistical software packages.

5.2.1.3 Approximate Joint and Marginal Densities

Having discussed obtaining the $\hat{\gamma}_i$'s, we now derive the general form of the approximate marginalization for mixed effects models. Our goal is to approximate $g(\boldsymbol{\gamma})$ using Taylor's method, and thus to approximate the entire joint pdf $f(\mathbf{Y}, \boldsymbol{\gamma}|\theta)$. Observe that since we've assumed $(Y_i, \gamma_i) \perp\!\!\!\perp (Y_{i'}, \gamma_{i'})$ when $i \neq i'$, we can write:

$$
\begin{aligned}
f(\mathbf{Y}, \boldsymbol{\gamma} \mid \theta) &= \prod_{i=1}^{k} f(Y_i, \gamma_i \mid \theta) \\
&= \prod_{i=1}^{k} f(Y_i \mid \gamma_i, \theta) f(\gamma_i \mid \theta) \\
&= \prod_{i=1}^{k} h(Y_i) \exp\left(g(\gamma_i)\right) (2\pi)^{-q/2} |\Sigma_i|^{-1/2} \left(-\frac{1}{2} (\gamma_i - \psi_i)^T \Sigma_i^{-1} (\gamma_i - \psi_i) \right) \\
&= \prod_{i=1}^{k} h(Y_i) \exp\left(g(\hat{\gamma}_i) + \dot{g}(\hat{\gamma}_i)(\gamma_i - \hat{\gamma}_i) + \frac{1}{2} (\gamma_i - \hat{\gamma}_i)^T \ddot{g}(\hat{\gamma}_i)(\gamma_i - \hat{\gamma}_i) + r(g, \gamma_i, \hat{\gamma}_i) \right) \\
&\qquad \times (2\pi)^{-q/2} |\Sigma_i|^{-1/2} \exp\left(-\frac{1}{2} (\gamma_i - \psi_i)^T \Sigma_i^{-1} (\gamma_i - \psi_i) \right)
\end{aligned}
$$

Where $\hat{\gamma}_i$ is the pivot point for the Taylor approximation and $r(g, \gamma_i, \hat{\gamma}_i)$ is the remainder term after a second-order Taylor approximation to $g(\gamma_i)$. Then, we can apply Proposition 3.1, the

matrix complete-the-square formula, to combine the quadratic terms for γ_i. Applying this formula to our expression for $f(\mathbf{Y}, \boldsymbol{\gamma} \mid \theta)$ gives:

$$\begin{aligned} f(\mathbf{Y}, \boldsymbol{\gamma}|\theta) = \prod_{i=1}^{k} \frac{h(Y_i)}{(2\pi)^{q/2}} |\Sigma_i|^{-1/2} \exp\left(g(\hat{\gamma}_i) + \dot{g}(\hat{\gamma}_i)(\gamma_i - \hat{\gamma}_i) + r(g, \gamma_i, \hat{\gamma}_i)\right) \\ \times \exp\left(-\frac{1}{2}(\hat{\gamma}_i - \psi_i)^T(-\ddot{g}(\hat{\gamma}_i))\left(\Sigma_i^{-1} - \ddot{g}(\hat{\gamma}_i)\right)^{-1}\Sigma_i^{-1}(\hat{\gamma}_i - \psi_i)\right) \\ \times \exp\left(-\frac{1}{2}(\gamma_i - \gamma_i^*)^T\left(\Sigma_i^{-1} - \ddot{g}(\hat{\gamma}_i)\right)(\gamma_i - \gamma_i^*)\right) \end{aligned} \tag{5.4}$$

We highlight two considerations from Equation (5.4) before moving forward.

First, note that the product $-\ddot{g}(\hat{\gamma}_i)\left(\Sigma_i^{-1} - \ddot{g}(\hat{\gamma}_i)\right)^{-1}\Sigma_i^{-1}$ simplifies to $-\ddot{g}(\hat{\gamma}_i)\left(I_q - \Sigma_i\ddot{g}(\hat{\gamma}_i)\right)^{-1}$. Additionally, if $\ddot{g}(\hat{\gamma}_i)$ is invertible, this simplifies further to $(\Sigma_i - \ddot{g}(\hat{\gamma}_i)^{-1})^{-1}$. The first simplification can be useful for numerical calculations, since it only involves one matrix inversion. The second simplification helps us understand better what this term represents. This is (the inverse of) the sum of the covariance matrix for γ_i and—when $\hat{\gamma}_i$ is considered as an MLE to the regression parameters from the offset GLM, as described in Section 5.2.1.1—the asymptotic variance estimate for $\hat{\gamma}_i$.

Finally, recognize that the final line of Equation (5.4) is the kernel of a multivariate normal distribution for γ_i with covariance matrix $\left(\Sigma_i^{-1} - \ddot{g}(\hat{\gamma}_i)\right)^{-1}$ and mean vector

$$\gamma_i^* = -\ddot{g}(\hat{\gamma}_i)\left(\Sigma_i^{-1} - \ddot{g}(\hat{\gamma}_i)\right)^{-1}\hat{\gamma}_i + \left(I_q - \ddot{g}(\hat{\gamma}_i)\Sigma_i\right)^{-1}\psi_i.$$

Then define:

$$\hat{f}(\gamma_i \mid Y_i, \theta) = (2\pi)^{-q/2}|\Sigma_i^{-1} - \ddot{g}(\hat{\gamma}_i)|^{1/2} \exp\left(-\frac{1}{2}(\gamma_i - \gamma_i^*)^T \left(\Sigma_i^{-1} - \ddot{g}(\hat{\gamma}_i)\right)(\gamma_i - \gamma_i^*)\right)$$

$$\hat{f}(Y_i \mid \theta) = h(Y_i)|\Sigma_i|^{-1/2}|\Sigma_i^{-1} - \ddot{g}(\hat{\gamma}_i)|^{-1/2} \exp\left(g(\hat{\gamma}_i) + \frac{1}{2}(\hat{\gamma}_i - \psi_i)^T \ddot{g}(\hat{\gamma}_i)\left(I_q - \Sigma_i \ddot{g}(\hat{\gamma}_i)\right)^{-1}(\hat{\gamma}_i - \psi_i)\right)$$

$$\hat{f}(Y_i, \gamma_i \mid \theta) = \hat{f}(\gamma_i | Y_i, \theta)\hat{f}(Y_i | \theta)$$

and observe that

$$f(\mathbf{Y}, \boldsymbol{\gamma} \mid \theta) = \prod_{i=1}^{k} \hat{f}(\gamma_i \mid Y_i, \theta)\hat{f}(Y_i \mid \theta) \exp\left(\dot{g}(\hat{\gamma}_i)(\gamma_i - \hat{\gamma}_i) + r(\gamma_i, \hat{\gamma}_i)\right).$$

Then clearly, since $\hat{f}(\gamma_i \mid Y_i, \theta)$ has the form of a proper density, if $\dot{g}(\hat{\gamma}_i)(\gamma_i - \hat{\gamma}_i) + r(\gamma_i, \hat{\gamma}_i)$ is zero or near-zero, $\hat{f}(Y_i \mid \theta)$ can act as an approximation to the difficult-to-find true marginal density $f(Y_i|\theta)$.

5.2.2 The Binomial and Poisson Approximations

In this subsection, we give precise specifications of key results from Section 5.2.1 as they pertain to the binomial and Poisson GLMMs.

5.2.2.1 Binomial Data

In the binomial model, we follow the same notation and assumptions outlined above. Critically, assume that clusters are independent, but observations within a cluster share a de-

pendence structure modeled by the random effects terms $\mathbf{Z}_i\gamma_i$. Let

$$
\begin{aligned}
y_{ij} &\sim \text{Bin}\left(m_{ij}, p_{ij}\right), \\
p_{ij} &= \frac{\exp\left(X_{ij}\beta + Z_{ij}\gamma_i\right)}{1+\exp\left(X_{ij}\beta + Z_{ij}\gamma_i\right)}, \\
\gamma_i &\sim \text{N}_q(\psi_i, \Sigma_i).
\end{aligned}
$$

Then we can write the principal elements of our approximation as

$$
\begin{aligned}
h(Y_i) &= \prod_{j=1}^{n} \binom{m_{ij}}{y_{ij}}, \\
g(\gamma_i) &= \sum_{j=1}^{n} \left(y_{ij}(X_{ij}\beta + Z_{ij}\gamma_i) - m_{ij}\log\left(1+\exp\left(X_{ij}\beta + Z_{ij}\gamma_i\right)\right)\right), \\
\ddot{g}(\gamma_i) &= -\mathbf{Z}_i^T V_{B_i} \mathbf{Z}_i,
\end{aligned}
$$

where V_{B_i} is the diagonal matrix of binomial variances for each observation, $m_{ij}p_{ij}(1-p_{ij})$. With t an $n\times 1$ vector, we use the matrix notation $D(t)$ to refer to the $n\times n$ diagonal matrix with the elements of t on the diagonal. Then $V_{B_i} = D(M_i)D(P_i)(I_n - D(P_i))$, where $M_i = \{m_{i1}, ..., m_{in}\}$ and $P_i = \{p_{i1}, ..., p_{in}\}$.

This results in an approximate marginal pdf for Y_i given θ in the Binomial case of

$$
\begin{aligned}
\hat{f}(Y_i|\theta) &= h(Y_i)|\Sigma_i|^{-1/2}|\Sigma_i^{-1} - \ddot{g}(\hat{\gamma}_i)|^{-1/2} \exp\left(g(\hat{\gamma}_i) + \frac{1}{2}(\hat{\gamma}_i - \psi_i)^T - \ddot{g}(\hat{\gamma}_i)\left(I_q - \Sigma_i\ddot{g}(\hat{\gamma}_i)\right)^{-1}(\hat{\gamma}_i - \psi_i)\right) \\
&= \left(\prod_{j=1}^{n}\binom{m_{ij}}{y_{ij}}\right)|\Sigma_i|^{-1/2}\left|\Sigma_i^{-1} + \mathbf{Z}_i^T\hat{V}_{B_i}\mathbf{Z}_i\right|^{-1/2} \\
&\qquad \times \exp\left(\sum_{j=1}^{n}\left(y_{ij}(X_{ij}\beta + Z_{ij}\hat{\gamma}_i) - m_{ij}\log\left(1+\exp\left(X_{ij}\beta + Z_{ij}\hat{\gamma}_i\right)\right)\right)\right) \\
&\qquad \times \exp\left(-\frac{1}{2}(\hat{\gamma}_i - \psi_i)^T\mathbf{Z}_i^T\hat{V}_{B_i}\mathbf{Z}_i\left(I_q + \Sigma_i\mathbf{Z}_i^T\hat{V}_{B_i}\mathbf{Z}_i\right)^{-1}(\hat{\gamma}_i - \psi_i)\right),
\end{aligned}
$$

where $\hat{V}_{B_i} = D(M_i)D(\hat{P}_i)(I_n - D(\hat{P}_i))$ and

$$\hat{P}_i = \{\hat{p}_{i1}, ..., \hat{p}_{in}\} = \left\{ \frac{\exp\left(X_{i1}\beta + Z_{i1}\hat{\gamma}_i\right)}{1 + \exp\left(X_{i1}\beta + Z_{i1}\hat{\gamma}_i\right)}, ..., \frac{\exp\left(X_{in}\beta + Z_{in}\hat{\gamma}_i\right)}{1 + \exp\left(X_{in}\beta + Z_{in}\hat{\gamma}_i\right)} \right\}.$$

5.2.2.2 Poisson Data

Next consider the Poisson model. We use the above notation and assume independence between clusters and structured dependence within clusters. Let

$$\begin{aligned} y_{ij} &\sim \text{Pois}\left(m_{ij}\lambda_{ij}\right), \\ \lambda_{ij} &= \exp\left(X_{ij}\beta + Z_{ij}\gamma_i\right), \\ \gamma_i &\sim \text{N}_q(\psi_i, \Sigma_i). \end{aligned}$$

Under the Poisson model, this gives

$$\begin{aligned} h(Y_i) &= \prod_{j=1}^{n} \frac{m_{ij}^{y_{ij}}}{y_{ij}!}, \\ g(\gamma_i) &= \sum_{j=1}^{n} \left(y_{ij}(X_{ij}\beta + Z_{ij}\gamma_i) - m_{ij}\exp(X_{ij}\beta + Z_{ij}\gamma_i) \right), \\ \ddot{g}(\gamma_i) &= -\mathbf{Z}_i^T V_{P_i} \mathbf{Z}_i, \end{aligned}$$

where $V_{P_i} = D(M_i)D(\lambda_i)$. M_i is defined as for the binomial, and $\lambda_i = \{\lambda_{i1}, ..., \lambda_{in}\}$.

Then we have an approximate marginal pdf for Y_i given θ in the Poisson case,

$$\begin{aligned}
\hat{f}(Y_i|\theta) &= h(Y_i)|\Sigma_i|^{-1/2}|\Sigma_i^{-1} - \ddot{g}(\hat{\gamma}_i)|^{-1/2} \exp\left(g(\hat{\gamma}_i) + \frac{1}{2}(\hat{\gamma}_i - \psi_i)^T \ddot{g}(\hat{\gamma}_i)\left(I_q - \Sigma_i \ddot{g}(\hat{\gamma}_i)\right)^{-1}(\hat{\gamma}_i - \psi_i)\right) \\
&= \left(\prod_{j=1}^{n} \frac{m_{ij}^{y_{ij}}}{y_{ij}!}\right) |\Sigma_i|^{-1/2} \left|\Sigma_i^{-1} + \mathbf{Z}_i^T \hat{V}_{P_i} \mathbf{Z}_i\right|^{-1/2} \\
&\quad \times \exp\left(\sum_{j=1}^{n} \left(y_{ij}(X_{ij}\beta + Z_{ij}\hat{\gamma}_i) - m_{ij}\exp\left(X_{ij}\beta + Z_{ij}\hat{\gamma}_i\right)\right)\right) \\
&\quad \times \exp\left(-\frac{1}{2}(\hat{\gamma}_i - \psi_i)^T \mathbf{Z}_i^T \hat{V}_{P_i} \mathbf{Z}_i \left(I_q + \Sigma_i \mathbf{Z}_i^T \hat{V}_{P_i} \mathbf{Z}_i\right)^{-1}(\hat{\gamma}_i - \psi_i)\right),
\end{aligned}$$

where $\hat{V}_{P_i} = D(M_i)D(\hat{\lambda}_i)$ and

$$\hat{\lambda}_i = \{\hat{\lambda}_{i1}, ..., \hat{\lambda}_{in}\} = \{\exp\left(X_{i1}\beta + Z_{i1}\hat{\gamma}_i\right), ..., \exp\left(X_{in}\beta + Z_{in}\hat{\gamma}_i\right)\}.$$

5.3 Simulation Results

In this section, we present small-sample simulation results for our marginal approximation. We begin by describing our simulation procedures in detail. We then report results on approximation error for the joint distribution and for the marginal DIC calculations. Finally, we present results for a simulated variable selection task using a minimum DIC model selection criterion based on DIC_b, DIC_j, and DIC_b.

5.3.1 Description of Simulation Procedures

As in the previous chapter, we conducted two sets of simulations to test our method. These correspond directly to the methods detailed in the previous chapter: the first assesses marginalization accuracy in the general setting, when both NR and Taylor approximation methods are being used; and the second examines the behavior of the method in a variable

selection task like the one presented in Chapter 4. The simulation examples detailed below are for models of the form

$$
\begin{aligned}
y_{ij}|X_{ij},\gamma_i &\overset{\text{iid}}{\sim} \text{exponential family} \\
\mathrm{E}\big[y_{ij}\big] &= \mu_{ij} \\
q(\mu_{ij}) &= X_{ij}\beta + \gamma_i \\
\gamma_i &\overset{\text{iid}}{\sim} N(0,\sigma^2).
\end{aligned}
$$

This is the prototypical longitudinal GLMM with time-varying[2] covariates and a single cluster effect. In Chapter 6, we will discuss extensions of this work to look at how our approximation fares when multiple random effects are used in a model.

In the first simulation set, data were again generated under a range of choices for k (number of observational units) and n (number of observations per unit). For a given k, a set of γ^*'s were sampled from the quantiles of a standard normal distribution according to the scheme

$$
\gamma_i^* = \Phi^{-1}\left(\frac{2i-1}{2k}\right), \qquad i \in \{1,...,k\}.
$$

Under the binomial scenario, we used $\gamma_i = \gamma_i^*$; while under the Poisson scenario, $\gamma_i = \gamma_i^*/2$. This was done because responses from the Poisson model are more sensitive to perturbations in the linear term $X_{ij}\beta + Z_{ij}\gamma_i$ than responses from the binomial model due to the exponential link function.

The covariate matrix $\mathbf{X}$ consisted of two columns—the first a column of ones, and the second a vector of time-varying covariates. For each cluster, values of the second covariate were obtained using a random walk with random bias. Figure 5.1 displays time-varying covariates generated under our method with $n = 20$ observations for four example individuals.

[2]We use the term time-varying to mean any covariates that change between observations taken on the same cluster, though we acknowledge that such changes do not always involve a difference in time of observation.

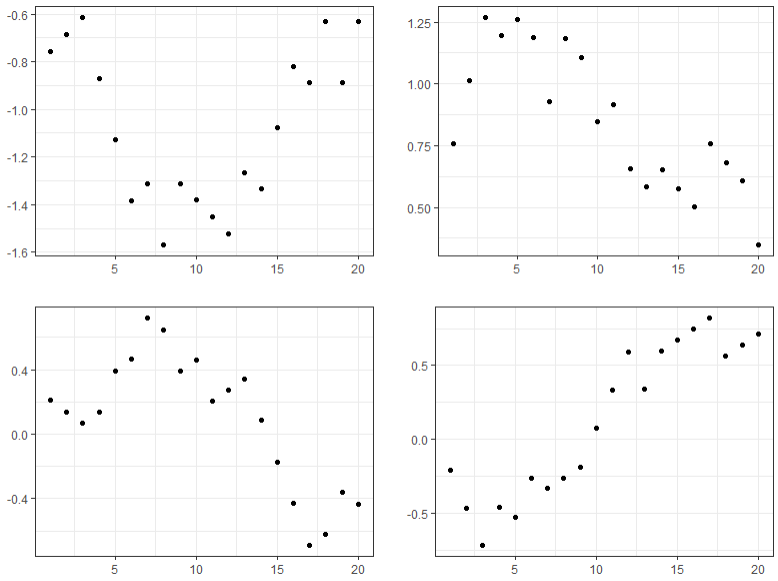

Figure 5.1: Example time-varying covariate data generated for four individuals with $n = 20$ observations, under the random walk simulation scheme described in this section.

Formally we denote as $x_{i,j,l}$ the l^{th} covariate measured on the j^{th} observation from the i^{th} cluster. Then $x_{i,j,1} \equiv 1 \, \forall \, i, j$, as described.

For each cluster i, an initial value of the second covariate $x_{i,1,2}$ was randomly selected from the quantiles of a $\text{N}(0, \sigma_0^2)$ distribution. Then a random walk was used to obtain the remaining $x_{i,1,2}$ values, where

$$x_{i,j,2} = x_{i,j-1,2} + e_{ij} \qquad\qquad j \in \{2, ..., n\},$$

where $e_{ij} \sim \text{N}(s_i, \sigma_e^2)$ are the steps of the random walk and $s_i \sim \text{N}(0, \sigma_s^2)$ is the random bias for cluster i. Values of s_i and e_{ij} were all chosen using the quantile sampling method and permuted into random order using a known random seed that changed for each simulation. Values of σ_0^2, σ_s^2, and σ_e^2 were chosen to give $x_{i,j,2}$ a total sample variance near 1.

We then generated response data, y. For Bernoulli simulations y_{ij} values were randomly sampled from the corresponding $\text{Bern}\left(p_{ij}\right)$ distributions, with $p_{ij} = \text{expit}\left(\beta_1 + x_{i,j,2}\beta_2 + \gamma_i\right)$ and

$(\beta_1, \beta_2) = (0, 1)$. For Poisson simulations y_{ij}'s were generated from $\text{Pois}(\lambda_{ij})$ distributions, with $\lambda_{ij} = \exp(\beta_1 + x_{i,j,2}\beta_2 + \gamma_i)$ and $(\beta_1, \beta_2) = (3, 1/4)$.

For each choice of k and n, data were simulated with random seeds selected by the authors for repeatability. A proportion of simulations failed due to extreme response data[3] and numerical instabilities in NR calculations. We present the number of successful simulations on which summary statistics are based in Table 5.1 below. Full results for these simulations are presented in Sections 5.3.2 and 5.3.3.

The second simulation set assesses automated model selection behavior. It is conducted in a similar fashion to the one from Chapter 4. We chose distinct (k, n) combinations based on what real-world data sets might look like; computational feasibility for repeated MCMC sampling; and our beliefs about how large n would need to be in order to obtain reasonable inferences for the model parameters. Automated model selection simulations were performed for five-trial binomial response data with $(k = 15, n = 20)$ and $(k = 20, n = 10)$; and for Poisson response data with $(k = 15, n = 10)$ and $(k = 20, n = 5)$.

For each simulation run, we created a design matrix including five covariates and an intercept: call them $\{x_1, x_2, x_3, x_4, x_5, x_6\}$. On the Binomial simulations, we again chose $\beta = [1/2 \quad 1 \quad 1/2 \quad 0 \quad 0 \quad 0]^T$; and on the Poisson simulations, we chose $\beta = [2 \quad 1/2 \quad 1/4 \quad 0 \quad 0 \quad 0]^T$. Covariate data were generated as a mix of cluster-level and observation-level variables. Specifically, x_1 was a column of 1's; x_2 and x_4 were cluster-level covariates chosen from quantiles of the $\text{N}(0, 1)$ distribution; and x_3, x_5, and x_6 were time-varying covariates generated using the profile method discussed above. For the binomial response data, γ_is were generated from a $\text{N}(0, 1)$ distribution, while under the Poisson scenario they came from a $\text{N}(0, 1/4)$ distribution. Response data were generated randomly according to Bernoulli and

[3]We refer to entire clusters where (binomial) every observation was a failure, or where every observation was a success, or (Poisson) where no events were observed. This problem becomes more severe as k increases, because there are more clusters in which extreme response data can occur.

Poisson distributions. As before, data were simulated with a new deterministic random seed in each simulation, to ensure that we obtained new-but-repeatable datasets each time.

Once the data were generated in each simulation, OpenBUGS was used to obtain posterior samples for θ and γ under the full model where all covariates were included, as well as a DIC_b score for that model. Based on these posterior samples, joint and marginal DICs were then computed in R.

We performed a backwards stepwise procedure using DIC_b as a selection criterion. Beginning with the full model, at each step DIC_b scores were obtained from OpenBUGS for the current model and for each model where one covariate was removed[4]. The model with the smallest DIC_b was selected as a new current model, the associated covariate removed, and this process was repeated. The stepwise procedure was stopped when either the DIC for all reduced models exceeded the DIC of the current model, or when we reached a model containing only the intercept term. We logged the order of removal, the final model selected, and posterior means and standard deviations on the β coefficients for the final model. Then we performed the same procedure, beginning again from the full model, using a numerical approximation to DIC_j as a selection criterion. Finally, we performed the same procedure with our approximation to DIC_m as a selection criterion. Results for the automated model selection simulation are presented in Section 5.3.4.

All Bernoulli simulations and the first set of Poisson simulations were performed on a single chain, each with a burn-in period of 1000 iterations and 5000 iterations included in the posterior sample. Chains were thinned by a factor of $\lceil n/10 \rceil$ (for Bernoulli) or n (for Poisson) to reduce autocorrelation. This was done because our DIC methods rely on performing computations on each of the posterior iterates, so having more posterior iterates during these computations incurs a greater computational cost. Thinning allows us to use fewer, more

[4]We only considered the removal of x_2 through x_6. The intercept analogue covariate, x_1, was forced into each model.

nearly independent iterates for these computations and improves our overall computational efficiency. We examined the convergence of three chains on a subset of the simulations we performed (2-3 combinations of k and n for each of the Bernoulli and Poisson schemes). Our thinning factor and number of burn-in iterations were chosen to ensure good convergence properties, based on the simulations examined.

All simulations were performed on a Linux server using OpenBUGS 3.2.3.

5.3.2 Pseudo-KL results

Table 5.1 displays pseudo-Kullback-Leibler information, as defined in the previous chapter, on the approximation accuracy of our method when time-varying covariates are included. Despite the addition of a new source of approximation error, the Newton-Rhapson root-finding step necessary to compute $\hat{\gamma}$, we see that pKL and relative pKL remain very small.

Bernoulli

k	n	m	pKL	$2pKL/\overline{D(\theta)}_j$	Sims
10	5	5	0.544	0.67%	43
	5	20	0.192	0.17%	95
	25	1	0.381	0.24%	31
	25	5	0.164	0.05%	94
	25	20	0.080	0.02%	95
30	5	5	1.352	0.57%	32
	5	20	0.492	0.14%	93
	25	1	1.670	0.36%	24
	25	5	0.429	0.04%	90
	25	20	0.170	0.01%	94

Poisson

k	n	pKL	$2pKL/\overline{D(\theta)}_j$	Sims
10	5	0.149	0.10%	11
	25	0.067	0.01%	11
30	5	0.291	0.06%	10
	25	0.127	0.01%	10

Table 5.1: Pseudo-Kullback-Leibler divergences between $f(\boldsymbol{\gamma}, \boldsymbol{\theta}|\mathbf{Y})$ and $\hat{f}(\boldsymbol{\gamma}, \boldsymbol{\theta}|\mathbf{Y})$, for simulated binomial and Poisson response data. The simulations column gives the number of successful simulations from which means of these values were computed.

Within the binomial simulations, we see that approximation error decreases as the number of observations per unit (n) increases. Similarly, as the number of binomial trials per observation

(m) increases, the approximation error decreases. Few simulations were successful with $n = 25$, $m = 1$ cases. This appears to relate to numerical singularities that occur if there exists $\hat{p}_{ij}$ very near zero or one. Essentially, this means that marginalization can fail when response data do not include enough information about both successes and failures for each individual.

In the binomial simulations, we do not see an appreciable change in pKL as k increases. The table shows that pKL increases in absolute terms, but relative pKL tends to remain similar except in the $n = 25$, $m = 1$ case where our simulation data are sparse.

The Poisson simulations we report have $m = 1$ only. This is because of the role m, the adjustment factor for the Poisson observational window size, plays in the likelihood equations. Whereas the shape of the distribution changes with m in the binomial response setting, in the Poisson setting m acts as a constant adjustment factor and is not of interest in examining approximation accuracy.

Among the Poisson simulations, we see similar results to those discussed above. Increasing n reduces pKL, in both absolute and relative terms. Increasing k increases pKL, but appears to have little effect on relative pKL. An extremely small number of Poisson simulations are unsuccessful compared with the binomial response simulations—computation only failed once, out of all the Poisson simulations we attempted in this simulation set.

5.3.3 Comparing the Approximation to Numerical Integration

For each of the scenarios presented in Table 5.1, Table 5.2 displays the average values of OpenBUGS DIC_b computations, our numerical approximations to DIC_j, and our approximating computation to DIC_m under both quadrature and our approximation method. Once again, both Gauss-Legendre and Gauss-Hermite quadrature were examined initially to assess

the number of nodes needed for convergence. Results presented below are for Gauss-Legendre quadrature, both for the binomial and the Poisson simulations. For the binomial simulations, 100 quadrature nodes were used for marginalization; and for the Poisson simulations, 200 were used. See the preceding chapter for further discussion of Gaussian quadrature and its application to GLMM marginalization.

Once again, we see that p_{Dm} and $\overline{D(\theta)}_m$ are very similar, whether computed using quadrature or using our approximation. And again, we see clear differences between these quantities, and BUGS and joint computations. There are surprising differences between computations for our approximation and for quadrature in the Binomial $n = 25$, $m = 20$ cases; but we note that these differences again appear to be due to a change in the quadrature calculations. Our approximations to p_{Dm} are relatively stable across all simulation designs, but quadrature approximations to p_{Dm} look very different in the Binomial $n = 25$, $m = 20$ simulations. We believe this is because of the choice of quadrature nodes, and because of the shape of the functions being numerically integrated. If not enough nodes are chosen and a function is very spiky, it is possible for the grid of quadrature nodes to miss the spike and fail to include the most important part of the distribution in their weighted averaging. Note that we do not observe these quadrature-approximation differences in our calculations of $\overline{D(\theta)}$, indicating that they are specific to the calculation of $D(\hat{\theta})$, not the average of $D\left(\theta^{(s)}\right)$'s for the posterior iterates. Here, the $D(\hat{\theta})$ function specifically is failing to be well-covered by our quadrature grid.

Obviously this issue is correctable by taking a larger grid, but we leave these points of disagreement in our work for pedagogical purposes. The three main difficulties of quadrature marginalization are the time quadrature takes, the functional transformations necessary to use quadrature approaches, and picking a sufficiently dense grid of quadrature nodes. This final issue complicates the use of quadrature for GLMM marginalization because the number

			p_D				$\overline{D(\theta)}$				DIC			
k	n	m	BUGS	Joint	Quad	Appx	BUGS	Joint	Quad	Appx	BUGS	Joint	Quad	Appx
Binomial														
	5	5	11.4	12.0	2.4	2.2	135.2	163.0	152.5	151.0	146.6	175.0	155.0	153.3
	5	20	11.0	12.4	2.4	2.4	202.7	232.1	233.1	232.8	213.7	244.5	235.5	235.1
10	25	1	11.4	11.7	2.5	2.3	294.4	319.9	310.2	309.5	305.8	331.6	312.7	311.8
	25	5	11.1	12.5	2.5	2.5	648.7	678.2	681.1	680.9	659.8	690.7	683.6	683.4
	25	20	11.2	12.5	2.4	2.8	1010.8	1040.2	1058.2	1056.6	1022.0	1052.7	1061.1	1059.0
	5	5	31.2	32.6	2.9	2.6	394.8	476.9	444.1	439.9	426.0	509.5	447.0	442.5
	5	20	31.0	32.7	2.8	2.7	604.4	690.3	692.6	691.4	635.4	723.1	695.4	694.1
30	25	1	31.7	32.6	2.9	2.6	850.7	931.8	899.4	893.9	882.4	964.4	902.2	896.5
	25	5	31.0	32.8	2.9	2.8	1939.4	2025.9	2034.4	2033.4	1970.4	2058.7	2037.3	2036.2
	25	20	30.9	32.8	3.1	2.9	3020.3	3106.5	3157.7	3155.4	3051.2	3139.3	3160.8	3158.3
Poisson														
10	5	—	11.1	12.6	2.7	2.6	287.3	302.8	320.9	320.9	298.4	315.4	323.5	323.6
	25	—	11.0	12.6	2.6	2.6	1468.8	1484.2	1518.2	1518.2	1479.8	1496.9	1520.8	1520.8
30	5	—	30.9	32.7	2.9	2.9	875.6	919.0	973.7	973.5	906.5	951.7	976.6	976.3
	25	—	31.3	32.8	2.9	2.9	4375.9	4420.0	4521.9	4521.8	4407.2	4452.9	4524.8	4524.7

Table 5.2: Comparison of p_D's and DIC's for various calculation methods on binomial and Poisson GLMMs. For the binomial GLMM, random effects variance $\sigma^2 = 1$ and quadrature is based on a 100-node calculation. For the Poisson GLMM, random effects variance $\sigma^2 = {}^1/_4$ and quadrature is based on a 200-node calculation.

of nodes necessary for correct computation depends on the dataset being analyzed and cannot be chosen in an ad hoc manner.

Results relative to DIC_b and DIC_j computations are much as we saw in the preceding chapter. We recognize that p_D is smaller for the marginal computation methods than for either the BUGS or joint computations. There is no clear relationship between $\overline{D(\theta)}$ under the BUGS, joint, and marginal computation methods. We will need to turn our attention to a model selection task to better understand how these criteria behave relative to one another.

Bernoulli

k	n	m	BUGS	Joint	Q-100	Appx
10	5	5	7	1	342	17
	5	20	6	1	343	11
	25	1	75	1	378	70
	25	5	76	1	378	58
	25	20	76	2	379	53
30	5	5	51	2	1014	139
	5	20	51	2	1005	71
	25	1	663	3	1128	3167
	25	5	674	3	1124	1252
	25	20	683	4	1119	1039

Poisson

k	n	BUGS	Joint	Q-200	Appx
10	5	22	3	921	12
	25	891	11	983	60
30	5	1329	8	2745	68
	25	118343	33	2875	1002

Table 5.3: Runtimes (in seconds) for DIC calculations. Since joint, quadrature, and approximation calculations are based on existing MCMC samples, each of these columns reports the additional time necessary for these calculations after the BUGS runtime has already elapsed.

Table 5.3 reports the runtimes of the various DIC computation methods. The number of quadrature nodes used for the Bernoulli simulations were $B_n = 100$, and for the Poisson we used $B_n = 200$. The joint computations execute very quickly, requiring only one algebraic computation for each posterior iterate $\theta^{(s)}$. The quadrature computations take appreciably longer, requiring B_n computations per posterior iterate. In contrast to Table 4.4, we find that the approximate marginal computations take considerably longer than the joint computations here, although they are still shorter than the quadrature computations in nearly every case. The increased runtimes for the approximate marginal computations are caused by the addition of the Newton-Rhapson root-finding step, which must be performed for ev-

ery posterior iterate $\theta^{(s)}$. In most cases, this is still faster than performing quadrature's B_n computations per posterior iterate. Runtimes for the approximate marginal computations in the $k = 30, n = 25, m = 1$ case are especially long, however. These suggest issues in our implementation of the Newton-Rhapson algorithm, though more investigation is needed. We do not believe our implementation is especially efficient, and we expect there is room for future research on speeding up our marginal computations.

BUGS runtimes here tend to be shorter than quadrature computations in the binomial case, but longer in the Poisson case. We expect the long Poisson runtimes for BUGS to be an effect of our thinning of the posterior iterates to deal with very strong autocorrelation. We remind the reader that we choose to thin our posterior samples because all numerical approximation methods to DIC_j and DIC_m require us to perform computations involving each posterior iterate. The computational cost of approximating DIC on unthinned, highly autocorrelated posterior observations can be extreme. We, therefore, need to run many more iterations of a Gibbs sampler on the Poisson to reach our desired $B_t = 5000$ samples. Since thinning ratios are smaller for the Binomial, less iterations (and less computational time) is required to reach 5000 posterior samples.

We believe adopting a centering parametrization for these models, as discussed in Section 4.1.1, may result in posterior sampling that requires less thinning. This would reduce the BUGS runtimes and allow for further analysis of the behavior of the various numerical approximation methods for DIC in the Poisson setting.

As in the previous chapter, the results presented above make the case for why we consider our marginal approximation superior to an approach like Gaussian quadrature. We have provided a closed-form expression for an approximate marginal density of the GLMM. This approximate marginal density can be used in DIC_m computations and gives very similar results to those obtained with quadrature methods, while not requiring evaluation over a grid of points of unknown size. We also find that our approximation method generally take

considerably less time to run than quadrature, despite the inclusion of a Newton-Rhapson step to find $\hat{\gamma}$.

5.3.4 Simulated Stepwise Variable Selection

Results from our simulated stepwise variable selection task are presented in Table 5.3. The list of criteria we consider for assessing the results of the stepwise algorithm are the same as those used in the preceding chapter: proportion of times the correct model is chosen, proportion of times no omission errors are committed, average number of omission errors, average number of inclusion errors, and our good model rate (GMR).

As before, we see that across nearly every metric, model selection by DIC_m outperforms selection by DIC_b or DIC_j. An interesting point to note, however, is that the BUGS DIC performs consistently worse than even the joint-distribution DIC when considering mixed effects models with random subject effects and time-varying covariates. We are not sure what causes this discrepancy—as previously discussed, we are not able to examine the technical details of OpenBUGS's DIC calculation algorithm—but the rate at which stepwise-by-DIC_b commits both omission and inclusion errors is surprising, when compared to the alternatives we considered.

5.4 Discussion

In this chapter, we have discussed the difference between REO and non-REO GLMMs and explained why the approximate marginalization developed in Chapter 4 is inappropriate for use with non-REO GLMMs. We have shown how the addition of a Newton-Rhapson step, equivalent to IRLS estimation of a generalized linear model with offset, allows us to extend our method to non-REO GLMMs. Section 5.2 presents computational formulas for

Model Type	k, n	Simulations	Method	Pr(Correct)	Pr(No O-Errors)	O-Errors	I-Errors	GMR
Binomial	15, 20	236	BUGS	0.09	0.59	0.43	1.34	0.42
			Joint	0.34	0.95	0.05	0.94	0.73
			Marginal	0.35	0.95	0.05	0.89	0.75
	20, 10	34	BUGS	0.06	0.65	0.35	1.65	0.24
			Joint	0.35	0.88	0.12	0.91	0.68
			Marginal	0.44	0.91	0.09	0.79	0.71
Poisson	15, 10	37	BUGS	0.11	0.73	0.27	1.19	0.59
			Joint	0.38	0.92	0.08	0.81	0.73
			Marginal	0.38	0.92	0.08	0.81	0.76
	20, 5	53	BUGS	0.06	0.70	0.30	1.40	0.51
			Joint	0.43	0.98	0.02	0.77	0.81
			Marginal	0.51	1.00	0.00	0.64	0.85

Table 5.4: Comparison of backwards stepwise selection output for BUGS, joint, and marginal calculation methods of DIC. Binomial response data are all generated assuming five Bernoulli trials.

an approximate marginal density in the logistic and Poisson GLMM settings, as well as a general expression that can be used for other GLMMs. A simulation-based small sample analysis in a common setting, the longitudinal model with time-varying covariates, has been provided.

The next chapter will propose a number of extensions to this research that we plan to pursue after the dissertation. We believe the approximate marginalization method we've developed for DIC_m has considerable potential for broader application in the area of longitudinal and hierarchical modeling.

Chapter 6

Future Directions

This brings us to the conclusion of the dissertation. In Chapter 1, we introduced our philosophy of statistics and the major themes of the dissertation. Chapter 2 presented previously published material on an application of mixed modeling in environmental epidemiology, as well as methodological work we undertook for that project. Chapters 3 through 5 examined the use of the deviance information criterion (DIC) in mixed models. We began with a discussion of information criteria for model selection and proceeded to discuss how marginalization affects the DIC and why we think the marginal DIC, DIC_m, is the quantity that should be considered for selection in mixed models. In Chapter 3, we provided two equivalent marginal expressions for the linear mixed model, the second arising from an application of the complete-the-square method. In Chapter 4, we used Taylor's theorem to develop an approximation to the marginal density for generalized linear mixed models (GLMMs) with repeated exchangeable observations (REO) and proved that our approximation converges almost surely to the true marginal density. Chapter 5 extended this method to a larger class of GLMMs by showing how root-finding methods could be incorporated to obtain the necessary pivots for Taylor's theorem.

In this chapter, we discuss our plans for future research on these topics.

6.1 Future Research on Missing Data Methods

In our work with Dr. Ulrike Luderer examining the effect of environmental pollutants on the human menstrual cycle, discussed in Chapter 2, we were unable to examine one of the response variables of interest: ovulation. This is a particularly interesting missing data problem we intend to return to after the dissertation. Ovulation status in these data is assessed based on monthly hormone profiles. If an "LH surge"—a one-day-long spike in luteinizing hormone levels—is observed, we determine that ovulation occurred. If such a surge is not observed, the question of ovulation status is more complicated.

Failing to observe an LH surge can happen in two ways: (1) an LH surge did not occur during a given menstrual cycle, or (2) an LH surge occurred but data were not collected on the day it occurred. Few of the cycles in our dataset have LH data for every day[1] during that cycle. "Successes" (cycles in which ovulation occurred) in our dataset are clear. "Failures" (cycles in which ovulation did not occur) are hard to distinguish from cycles in which ovulation occurred but was not observed. This is an example of missing not at random (MNAR) data.

Ovulation status in these data can be viewed as trinomial response data, with individual cycles being classified as ovulatory, anovulatory, or unknown. Some cycles are clearly anovulatory: if data are available for every day of a menstrual cycle and no LH surge is detected, we can reasonably state that ovulation did not occur. Some cycles are clearly unknown: if data were not collected on any day in the cycle, we can know nothing about whether ovulation occurred. Most cycles where ovulation isn't observed are hard to classify: we have seen that the LH surge *did not* happen on many days during the cycle, but we are missing data on a few days and cannot conclusively state that the cycle is anovulatory.

[1] During the follicular phase. Refer back to Section 2.X for a discussion of data-gathering procedures.

Bayesian methods work well for data of this sort, allowing us to probabilistically model whether each cycle should be classified as ovulatory, anovulatory, or unknown. This approach will let us address the question of whether environmental pollutants affect the probability of ovulation in humans, an important contribution to the emerging literature on environmental effects on fertility.

6.2 Future Research on GLMM Marginalization

Below, we detail areas for continued research based on our approximate marginalization method for GLMMs. The first two areas deal with technical aspects of our method and its implementation in computer software. The latter two extend the work in this dissertation to more general models than we have yet been able to consider.

6.2.1 Properties of the Marginal Approximation

As we discuss briefly in Chapter 4, Cai and Dunson (2006) have also provided an approximation to the marginal distribution for GLMMs, using some of the same tools as us. It is not clear how their approximation compares to ours. We believe both should be similarly accurate, although ours provides a more concise approximate expression. We intend to further investigate how these two approximations compare, both mathematically and through simulation.

Further, as discussed in the appendix, we have proven a proposition about the integrability of our marginal approximation, $\hat{f}(y \mid \theta)$. This result, Proposition A.1, establishes that if a certain expectation related to the Taylor remainder exists, our approximations will be integrable. We know they are integrable at the limit because of Proposition 4.1, which establishes that our joint approximation converges almost surely to the true joint density $f(y, \gamma \mid \theta)$ in the

REO case, when $\hat{\gamma} \xrightarrow{\text{a.s.}} \gamma$. Thus far, Proposition A.1 has not been enough to allow us to state definitively that our small-sample marginal approximations are integrable—but if further research lets us establish this, we believe we can use our marginal approximation to define a closed-form approximate marginal distribution for REO GLMMs (where $\hat{\gamma}$ is an analytic function of the data), and possibly to define useful approximate marginal distributions for non-REO GLMMs as well. We also plan to prove the comparable result to Proposition 4.1 for the marginal approximation used in Chapter 5, where $\hat{\gamma}$ is obtained by approximation and convergence is more nuanced.

6.2.2 Methods for Achieving Greater Computational Efficiency

Simulations presented in Chapter 5 of this dissertation use a basic Newton-Rhapson function coded by the author based on convenience and not computational efficiency. Although the computational time required by our method was generally shorter than the quadrature alternative presented, we expect that this time can be substantially reduced. Two strategies for this are particularly promising.

The first is to use one-step or two-step Newton-Rhapson rather than the method we implemented, which iterated until either our iterated approximation to γ converged ($||\gamma^{[i]} - \gamma^{[i-1]}|| < 10^{-8}$) or 30 steps were completed without convergence. Our NR algorithm benefits from having good initial values in the Bayesian setting: when we calculate $\hat{\gamma}_{(s)}$ for some $\theta^{(s)}$ sampled from the posterior, it is reasonable to start our Newton-Rhapson algorithm from $\gamma^{(s)}$, the sampled posterior value associated with $\theta^{(s)}$. Because we have good initial guesses through our posterior sample, it may be possible to use only one or two steps of the NR algorithm and still be accurate. A one-step or two-step algorithm would greatly reduce the computational burden of the method, and may still result in $\hat{\gamma}$ satisfying $\dot{g}(\hat{\gamma}) \doteq 0$, close enough that our approximation works well. Other initial values will also be investigated.

The second promising strategy, discussed in Section 5.2.1.2 but not yet implemented for our simulations, is to use the equivalence of our root-finding problem and existing methods for maximum likelihood estimation in a generalized linear model (GLM) with an $X\beta$ offset. There are two major advantages to this strategy. First, it takes advantage of well developed computational methods for improving GLM-fitting. Second, it simplifies obtaining the $\hat{\gamma}$'s for our non-REO GLMM approximation. If these values can be obtained existing statistical software, our method will be easier for broad scientific use.

6.2.3 Small-Sample Results When $q > 1$

In Chapter 5, we developed an approximate marginalization method for GLMMs that permits multiple random effects—that is, the random effects vector for cluster i, γ_i, is a random vector of length q. To date, our simulation work (Sections 4.4 and 5.3) has only involved adding time-varying covariates, not multivariate random effects. This is the first intended extension of our method.

Multivariate random effects can take many forms. The two most common are random slope models and multi-level mixed effects models, where more than one type of clustering is present in the dataset. Approximate marginalization of random slope models is necessarily an extension of our non-REO GLMM method from Chapter 5, since random slopes presuppose time-varying covariates. Multi-level mixed effects models, however, may still permit us to use our REO GLMM marginalization method. We have begun working on this problem, but further investigation is necessary. Below, we provide some initial notes on the marginalization of nested multi-level random effects based on Thurmond et al. (2005)'s cow abortion dataset. This dataset assumes that the probability of a cow having an abortion (miscarriage) during a given pregnancy involves both a herd-level and a cow-level random effect.

Similar to before, let $\mathbf{Y} = \{Y_i : i \in 1, ..., I\}$ be our response data, consisting of Bernoulli random variables indicating whether a given pregnancy ended in abortion[2]. Here Y_i is a vector of such observations on the entire herd i. Within each herd we assume that there are J_i cows and we write the vector Y_{ij} to represent repeated responses in time on cow j in herd i. We assume that there are K_{ij} pregnancies on cow j in herd i, and that y_{ijk} denotes whether pregnancy k on cow j in herd i ended in abortion.

A simple model for these data can be written as

$$\begin{aligned} y_{ijk} &\sim \text{Bern}\left(p_{ij}\right) \\ \text{logit}\left(p_{ij}\right) &= \mu + \gamma_i + \eta_{ij} \\ \gamma_i &\overset{\text{iid}}{\sim} \text{N}\Big(0, \sigma_h^2\Big) \\ \eta_{ij} &\overset{\text{iid}}{\sim} \text{N}\Big(0, \sigma_c^2\Big), \end{aligned}$$

where γ_i is the random effect for herd i and η_{ij} is the random effect for cow j in herd i. We use θ to represent the parameter vector $\{\mu, \sigma_h^2, \sigma_c^2\}$. We use $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$ to represent the full vectors of herd and cow random effects respectively. We ignore time dependent covariates for this present development, though we will include them in the later research.

Then the joint density for these data, conditional on the parameters and the random effects, is

$$\begin{aligned} f(\mathbf{Y} \mid \boldsymbol{\gamma}, \boldsymbol{\eta}, \theta) &= \prod_{i=1}^{I} \prod_{j=1}^{J_i} \prod_{k=1}^{K_{ij}} \left[\left(\frac{\exp(\mu + \gamma_i + \eta_{ij})}{1 + \exp(\mu + \gamma_i + \eta_{ij})} \right)^{y_{ijk}} \left(\frac{1}{1 + \exp(\mu + \gamma_i + \eta_{ij})} \right)^{1 - y_{ijk}} \right] \\ &= \prod_{i=1}^{I} \prod_{j=1}^{J_i} \prod_{k=1}^{K_{ij}} \left[\exp(\mu + \gamma_i + \eta_{ij})^{y_{ijk}} \left(\frac{1}{1 + \exp(\mu + \gamma_i + \eta_{ij})} \right) \right] \\ &= \exp\left[\sum_{i=1}^{I} \sum_{j=1}^{J_i} \sum_{k=1}^{K_{ij}} \left(y_{ijk}(\mu + \gamma_i + \eta_{ij}) - \log\left(1 + \exp(\mu + \gamma_i + \eta_{ij})\right) \right) \right]. \end{aligned}$$

[2]We now adopt the convention that i runs from 1 to I, because we want to make k available for another level of indexing on our data. We similarly change index bounds for j.

For a given herd i, we write our $g(\cdot)$ function as

$$\begin{aligned} g(\gamma_i, \eta_{i1}, ..., \eta_{iJ_i}) &= \sum_{j=1}^{J_i} \sum_{k=1}^{K_{ij}} \left(y_{ijk}(\mu + \gamma_i + \eta_{ij}) - \log\left(1 + \exp(\mu + \gamma_i + \eta_{ij})\right)\right) \\ &= \sum_{j=1}^{J_i} \left(y_{ij\cdot}(\mu + \gamma_i + \eta_{ij}) - K_{ij} \log\left(1 + \exp(\mu + \gamma_i + \eta_{ij})\right)\right). \end{aligned}$$

where $y_{ij\cdot} = \sum_k y_{ijk}$.

In order to use our approximation to the marginal density, then, we need to find $\hat{\gamma}_i, \hat{\eta}_{i1}, ..., \hat{\eta}_{iJ_i}$ satisfying $\dot{g}(\hat{\gamma}_i, \hat{\eta}_{i1}, ..., \hat{\eta}_{iJ_i}) = 0$. This means solving the for the simultaneous roots of the system of equations

$$\begin{aligned} \frac{\partial}{\partial \gamma_i} g(\gamma_i, \eta_{i1}, ..., \eta_{iJ_i}) &= y_{i\cdot\cdot} - \sum_{j=1}^{J_i} K_{ij} \frac{\exp(\mu + \gamma_i + \eta_{ij})}{1 + \exp(\mu + \gamma_i + \eta_{ij})} = 0 \\ \frac{\partial}{\partial \eta_{i1}} g(\gamma_i, \eta_{i1}, ..., \eta_{iJ_i}) &= y_{i1\cdot} - K_{i1} \frac{\exp(\mu + \gamma_i + \eta_{i1})}{1 + \exp(\mu + \gamma_i + \eta_{i1})} = 0 \\ &\vdots \\ \frac{\partial}{\partial \eta_{iJ_i}} g(\gamma_i, \eta_{i1}, ..., \eta_{iJ_i}) &= y_{iJ_i\cdot} - K_{iJ_i} \frac{\exp(\mu + \gamma_i + \eta_{iJ_i})}{1 + \exp(\mu + \gamma_i + \eta_{iJ_i})} = 0. \end{aligned}$$

We show below that a set of simultaneous roots are obtained with the expressions

$$\hat{\gamma}_i = c - \mu \qquad \text{and} \qquad \hat{\eta}_{ij} = \log\left(\frac{y_{ij\cdot}}{K_{ij} - y_{ij\cdot}}\right) - c,$$

where c is an arbitrary constant. Note that $\log\left({}^{y_{ij\cdot}}/_{(K_{ij} - y_{ij\cdot})}\right)$ is the sample log odds for abortions in cow j of herd i. We prefer to choose

$$c = \log\left(\frac{y_{i\cdot\cdot}}{J_i K_i - y_{i\cdot\cdot}}\right),$$

because this is our usual choice for $\hat{\gamma}_i$ when the cow-level random effects, η_{ij}, are ignored. This is the sample log odds for abortions in all cows of herd i.

Then $\frac{\partial}{\partial \gamma_i} g(\hat{\gamma}_i, \hat{\eta}_{i1}, ..., \hat{\eta}_{iJ_i})$ is equal to zero because

$$\begin{aligned}
\frac{\partial}{\partial \gamma_i} g(\hat{\gamma}_i, \hat{\eta}_{i1}, ..., \hat{\eta}_{iJ_i}) &= y_{i\cdot\cdot} - \sum_{j=1}^{J_i} K_{ij} \frac{\exp(\mu + \hat{\gamma}_i + \hat{\eta}_{ij})}{1 + \exp(\mu + \hat{\gamma}_i + \hat{\eta}_{ij})} \\
&= y_{i\cdot\cdot} - \sum_{j=1}^{J_i} K_{ij} \frac{\exp\left(\log\left(\frac{y_{ij\cdot}}{K_{ij} - y_{ij\cdot}}\right)\right)}{1 + \exp\left(\log\left(\frac{y_{ij\cdot}}{K_{ij} - y_{ij\cdot}}\right)\right)} \\
&= y_{i\cdot\cdot} - \sum_{j=1}^{J_i} K_{ij} \frac{\left(\frac{y_{ij\cdot}}{K_{ij} - y_{ij\cdot}}\right)}{1 + \left(\frac{y_{ij\cdot}}{K_{ij} - y_{ij\cdot}}\right)} \\
&= y_{i\cdot\cdot} - \sum_{j=1}^{J_i} K_{ij} \frac{\left(\frac{y_{ij\cdot}}{K_{ij} - y_{ij\cdot}}\right)}{\left(\frac{K_{ij}}{K_{ij} - y_{ij\cdot}}\right)} \\
&= y_{i\cdot\cdot} - \sum_{j=1}^{J_i} K_{ij} \left(\frac{y_{ij\cdot}}{K_{ij}}\right) \\
&= y_{i\cdot\cdot} - \sum_{j=1}^{J_i} y_{ij\cdot} \\
&= 0.
\end{aligned}$$

And for each j, $\frac{\partial}{\partial \eta_{ij}} g(\hat{\gamma}_i, \hat{\eta}_{i1}, ..., \hat{\eta}_{iJ_i})$ is also equal to zero,

$$\begin{aligned}
\frac{\partial}{\partial \eta_{ij}} g(\hat{\gamma}_i, \hat{\eta}_{i1}, ..., \hat{\eta}_{iJ_i}) &= y_{ij\cdot} - K_{ij} \frac{\exp(\mu + \hat{\gamma}_i + \hat{\eta}_{ij})}{1 + \exp(\mu + \hat{\gamma}_i + \hat{\eta}_{ij})} \\
&= y_{ij\cdot} - K_{ij} \frac{\exp\left(\log\left(\frac{y_{ij\cdot}}{K_{ij} - y_{ij\cdot}}\right)\right)}{1 + \exp\left(\log\left(\frac{y_{ij\cdot}}{K_{ij} - y_{ij\cdot}}\right)\right)} \\
&= y_{ij\cdot} - K_{ij} \frac{\left(\frac{y_{ij\cdot}}{K_{ij} - y_{ij\cdot}}\right)}{1 + \left(\frac{y_{ij\cdot}}{K_{ij} - y_{ij\cdot}}\right)} \\
&= y_{ij\cdot} - K_{ij} \left(\frac{y_{ij\cdot}}{K_{ij}}\right) \\
&= y_{ij\cdot} - y_{ij\cdot} \\
&= 0.
\end{aligned}$$

This gives us a choice of $\hat{\gamma}_i, \hat{\eta}_{i1}, ..., \hat{\eta}_{iJ_i}$ satisfying $\dot{g}(\hat{\gamma}_i, \hat{\eta}_{i1}, ..., \hat{\eta}_{iJ_i}) = 0$, which means we are able to use Chapter 4's REO GLMM approximation to the marginal. Note, however, that there are infinitely many such solutions, based on one's choice of c above. We do not yet know how the choice of c affects our approximation.

Further work is needed with this problem, and with the approximate marginalization of other multivariate mixed effects models. As we stated above, the non-REO marginalization method we developed in Chapter 5 applies when $q > 1$; and our asymptotic convergence result, Proposition 4.1, applies when $q > 1$ as well, as in the herds-and-cows example above. Simulating these scenarios to look at small-sample convergence properties and elaborating the REO method to handle multi-level REO situations like the one described above are our first priorities, moving forward.

6.2.4 More Random Effects Distributions

Another extension, which can be applied to both the REO and non-REO approximations, is to expand the classes of random effects distributions that can be used with our approximation to the marginal. Our marginal approximation relies on the normality of the random effects to combine the random effects density, $P(\gamma \mid \theta)$ with the conditional density for the data, $f(y \mid \gamma, \theta)$. The Taylor approximation we introduce in Chapters 4 and 5 allows us to use the complete-the-square formula to algebraically isolate γ in a normal kernel for easy integration.

Subject to "the devil is in the details," our method will in theory extend to GLMMs where random effects are modeled with a mixture-of-normals distribution. Random effects following a mixture-of-normals distribution can be handled similarly to the method we developed, resulting in an approximate conditional density for γ, $\hat{f}(\gamma \mid y, \theta)$ that will also be a mixture of normals. As long as $\hat{f}(\gamma \mid y, \theta)$ is analytically obtainable and a fully specified density function, our marginal approximation method should stay relatively unchanged.

Extending our method to handle random effects with mixture-of-normals distributions also opens the way for us to look at nonparametric random effects distributions. If we can show our method works for mixture-of-normals random effects, it should be straightforward to show that it works for random effects distributed as Dirichlet process mixtures (DPMs) of normals. This raises a secondary question of whether it is sensible to consider DIC for model selection in models like the GLMM with DPM random effects distributions, which we will duly investigate. Irrespective of this, however, we believe an approximate marginal form for DPM GLMMs is of interest in and of itself.

Bibliography

Akaike, Hirotugu (1974). "A new look at the statistical model identification". In: *IEEE Transactions on Automatic Control* AC-19.6, pp. 716–723. DOI: 10.1109/TAC.1974.1100705.

Ando, Tomohiro (2007). "Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models". In: *Biometrika* 94.2, pp. 443–458. DOI: 10.1093/biomet/asm017.

Aquilina, Noel J., Juana Mari Delgado-Saborit, Claire Meddings, Stephen Baker, et al. (2010). "Environmental and biological monitoring of exposures to PAHs and ETS in the general population". In: *Environment International* 36.7, pp. 763–771. DOI: 10.1016/j.envint.2010.05.015.

ATSDR (1995). *Agency for Toxic Substances and Disease Registry. Toxicological Profile for polycyclic aromatic hydrocarbons*. Atlanta, GA: US Department of Health and Human Services, Public Health Service.

Borman, S.M., P.J. Christian, I.G. Sipes, and P.B. Hoyer (2000). "Ovotoxicity in female Fischer rats and B6 mice induced by low-dose exposure to three polycyclic aromatic hydrocarbons: comparison through calculation of an ovotoxic index". In: *Toxicology and Applied Pharmacology* 167.3, pp. 191–198. DOI: 10.1006/taap.2000.9006.

Box, George E.P. and Norman R. Draper (1987). *Empirical Model Building and Response Surfaces*. New York, NY: John Wiley and Sons.

Breslow, N.E. (1984). "Extra-poisson variation in log-linear models". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 33.1, pp. 38–44.

Breslow, N.E. and D.G. Clayton (1993). "Approximate inference in generalized linear mixed models". In: *Journal of the American Statistical Association* 88.421, pp. 9–25. DOI: 10.1080/01621459.1993.10594284.

Brooks, S.J. (2002). "Discussion of 'Bayesian measures of model complexity and fit'". In: *Journal of the Royal Statistical Society: Series B* 24.4, pp. 616–618. DOI: 10.1111/1467-9868.00353.

Cai, Bo and David B. Dunson (2006). "Bayesian covariance selection in generalized linear mixed models". In: *Biometrics* 62.2, pp. 446–457. DOI: 10.1111/j.1541-0420.2005.00499.x.

Cavanaugh, Joseph E. (1997). "Unifying the derivations for the Akaike and corrected Akaike information criteria". In: *Statistics & Probability Letters* 33.2, pp. 201–208. DOI: 10.1016/S0167-7152(96)00128-9.

Cavanaugh, Joseph E. and Andrew A. Neath (1999). "Generalizing the derivation of the Schwarz information criterion". In: *Communications in Statistics – Theory and Methods* 28.1, pp. 49–66. DOI: 10.1080/03610929908832282.

Celeux, G., F. Forbes, C.P. Robert, and D.M. Titterington (2006). "Deviance information criteria for missing data models". In: *Bayesian Analysis* 1.4, pp. 651–674. DOI: 10.1214/06-BA122.

Çetin, Mustafa S., Fletcher Christensen, Christopher C. Abbott, Julia M. Stephen, et al. (2014). "Thalamus and posterior temporal lobe show greater inter-network connectivity at rest and across sensory paradigms in schizophrenia". In: *NeuroImage* 97, pp. 117–126. DOI: 10.1016/j.neuroimage.2014.04.009.

Chen, Changzhong, Xiaobin Wang, Lihua Wang, Fan Yang, et al. (2005). "Effect of Environmental Tobacco Smoke on Levels of Urinary Hormone Markers". In: *Environmental Health Perspectives* 113.4, pp. 412–417. DOI: 10.1289/ehp.7436.

Christensen, Ronald R. (1997). *Log-Linear Models and Logistic Regression*. 2nd. New York, NY: Springer.

— (2017). "Plane Answers to Complex Questions: the Theory of Linear Models, 5th ed". Personal correpondence.

Christensen, Ronald R. and Fletcher G.W. Christensen (2009). "Letters to the editor". In: *The American Statistician* 63.2, p. 197. DOI: 10.1198/tast.2009.0037.

Christensen, Ronald R., Wesley O. Johnson, Adam Branscum, and Timothy E. Hanson (2010). *Bayesian Ideas and Data Analysis*. Boca Raton, FL: CRC Press.

Crouch, Edmund A.C. and Donna Spiegelman (1990). "The evaluation of integrals of the form $\int_{-\infty}^{+\infty} f(t)\exp(-t^2)dt$: application to logistic-normal models". In: *Journal of the American Statistical Association* 85.410, pp. 464–469. DOI: 10.1080/01621459.1990.10476222.

Daniels, M.J. and J.W. Hogan (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Boca Raton, FL: CRC Press.

Ferguson, Thomas S. (1996). *A Course in Large Sample Theory*. Boca Raton, FL: CRC Press.

Geisser, Seymour and William F. Eddy (1979). "A predictive approach to model selection". In: *Journal of the American Statistical Association* 74.365, pp. 153–160. DOI: 10.1080/01621459.1979.10481632.

Gelfand, A.E. and D.K. Dey (1994). "Bayesian model choice: asymptotics and exact calculations". In: *Journal of the Royal Statistical Society: Series B* 56.3, pp. 501–514.

Gelfand, Alan E., Sujit K. Sahu, and Bradley P. Carlin (1996). "Efficient parametrisations for generalized linear mixed models". In: *Bayesian Statistics 5*. Ed. by J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith. Oxford, UK: Oxford University Press, pp. 227–246.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, et al. (2013). *Bayesian Data Analysis*. 3rd. Boca Raton, FL: CRC Press.

Hammersley, J.M. and D.C. Handscomb (1964). *Monte Carlo Methods*. New York, NY: John Wiley and Sons.

Harlow, Bernard L. and Lisa B. Signorello (2000). "Factors associated with early menopause". In: *Maturitas* 35.1, pp. 3–9. DOI: 10.1016/S0378-5122(00)00092-X.

Help, AdSense (2017). *Ad targeting: how ads are targeted to your site*. Google Inc. URL: https://support.google.com/adsense/answer/9713?hl=en (visited on).

Hill, R.H., S.L. Head, S. Baker, M. Gregg, et al. (1995). "Pesticide residues in urine of adults living in the United States: reference range concentrations". In: *Environmental Research* 71.2, pp. 99–108. DOI: 10.1006/enrs.1995.1071.

Hinde, John (1982). "Compound Poisson regression models". In: *GLIM 82: Proceedings of the International Conference on Generalized Linear Models*. Ed. by Robert Gilchrist. New York, NY: Springer-Verlag, pp. 109–121.

Hodges, James S. and Daniel J. Sargent (2001). "Counting degrees of freedom in hierarchical and other richly-parameterized models". In: *Biometrika* 88.2, pp. 367–379. DOI: 10.1093/biomet/88.2.367.

IARC (1983). *International Agency for Research on Cancer. Polynuclear aromatic compounds, part 1, chemical, environmental and experimental data*. Lyons, France: World Health Organization.

— (2010). *International Agency for Research on Cancer. Some non-heterocyclic polycyclic aromatic hydrocarbons and some related exposures*. Lyons, France: World Health Organization.

Kline, Morris (1972). *Mathematical Thought from Ancient to Modern Times, Vol. 2*. New York, NY: Oxford University Press.

Kullback, Solomon and Richard A. Leibler (1951). "On information and sufficiency". In: *Annals of Mathematical Statistics* 22.1, pp. 79–86. DOI: 10.1214/aoms/1177729694.

Li, Zheng, Courtney D. Sandau, Lovisa C. Romanoff, Samuel P. Caudill, et al. (2008). "Concentration and profile of 22 urinary polycyclic aromatic hydrocarbon metabolites

in the US population". In: *Environmental Research* 107.3, pp. 320–331. DOI: 10.1016/j.envres.2008.01.013.

Liang, Kung-Yee and Scott L. Zeger (1986). "Longitudinal data analysis using generalized linear models". In: *Biometrika* 73.1, pp. 13–22. DOI: 10.1093/biomet/73.1.13.

Little, Roderick J.A. and Donald B. Rubin (2002). *Statistical Analysis with Missing Data*. 2nd. Somerset, NJ: John Wiley and Sons.

Luderer, Ulrike, Fletcher Christensen, Wesley O. Johnson, Jianwen She, et al. (2017). "Associations between urinary biomarkers of polycyclic aromatic hydrocarbon exposure and reproductive function during menstrual cycles in women". In: *Environment International* 100, pp. 110–120. DOI: 10.1016/j.envint.2016.12.021.

Lunn, David, Christopher Jackson, Nicky Best, Andrew Thomas, and David Spiegelhalter (2013). *The BUGS Book*. Boca Raton, FL: CRC Press.

Mattison, Donald R. (1979). "Difference in sensitivity of rat and mouse primordial oocytes to destruction by polycyclic aromatic hydrocarbons". In: *Chemico-Biological Interactions* 28.1, pp. 133–137. DOI: 10.1016/0009-2797(79)90120-0.

Mattison, Donald R., David R. Plowchalk, M. Jane Meadows, Michael M. Miller, et al. (1989). "The effect of smoking on oogenesis, fertilization, and implantation". In: *Seminars in Reproductive Medicine* 7.4, pp. 291–304. DOI: 10.1055/s-2007-1021411.

Menzie, Charles A., Bonnie B. Potocki, and Joseph Santodonato (1992). "Exposure to carcinogenic PAHs in the environment". In: *Environmental Science and Technology* 26.7, pp. 1278–1284. DOI: 10.1021/es00031a002.

Neal, Michael S., Jiping Zhu, and Warren G. Foster (2008). "Quantification of benzo[a]pyrene and other PAHs in the serum and follicular fluid of smokers versus non-smokers". In: *Reproductive Toxicology* 25.1, pp. 100–106. DOI: 10.1016/j.reprotox.2007.10.012.

Neal, Michael S., Jiping Zhu, Alison C. Holloway, and Warren G. Foster (2007). "Follicle growth is inhibited by benzo-[a]-pyrene, at concentrations representative of human expo-

sure, in an isolated rat follicle culture assay". In: *Human Reproduction* 22.4, pp. 961–967. DOI: 10.1093/humrep/del487.

Nelder, J.A. and R.W.M. Wedderburn (1972). "Generalized linear models". In: *Journal of the Royal Statistical Society: Series A* 135.3, pp. 370–384. DOI: 10.2307/2344614.

NHANES (2009). *National Health and Nutrition Examination Survey (NHANES). Fourth national report on human exposure to environmental chemicals*. Atlanta, GA: Department of Health, Human Services, Centers for Disease Control, and Prevention.

Ogden, T.L. (2010). "Handling results below the level of detection". In: *Annals of Occupational Hygeine* 54.3, pp. 255–256. DOI: 10.1093/annhyg/mep099.

Overstall, Antony M. and Jonathan J. Forster (2010). "Default Bayesian model determination methods for generalised linear mixed models". In: *Computational Statistics and Data Analysis* 54.12, pp. 3269–3288. DOI: 10.1016/j.csda.2010.03.008.

Romanoff, Lovisa C., Zheng Li, Kisha J. Young, Nelson C. Blakely III, et al. (2006). "Automated solid-phase extraction method for measuring urinary polycyclic aromatic hydrocarbon metabolites in human biomonitoring using isotope-dilution gas chromatography high-resolution mass spectrometry". In: *Journal of Chromatography B* 835.1-2, pp. 47–54. DOI: 10.1016/j.jchromb.2006.03.004.

Sadeu, J.C. and Warren G. Foster (2011). "Effect of in vitro exposure to benzo[a]pyrene, a component of cigarette smoke, on folliculogenesis, steroidogenesis and oocyte nuclear maturation". In: *Reproductive Toxicology* 31.4, pp. 402–408. DOI: 10.1016/j.reprotox.2010.12.006.

Saefken, Benjamin, Thomas Kneib, Clara-Sophie van Waveren, and Sonja Greven (2014). "A unifying approach to the estimation of the conditional Akaike information in generalized linear mixed models". In: *Electronic Journal of Statistics* 8.1, pp. 201–225. DOI: 10.1214/14-EJS881.

Sahu, Sujit K. (2002). "Discussion of 'Bayesian measures of model complexity and fit'". In: *Journal of the Royal Statistical Society: Series B* 24.4, pp. 625–626. DOI: 10.1111/1467-9868.00353.

Schwarz, Gideon (1978). "Estimating the dimension of a model". In: *Annals of Statistics* 6.2, pp. 461–464. DOI: 10.1214/aos/1176344136.

Shopland, D.R., D.M. Burns, N.L. Benowitz, and R.H. Amacher (2001). "Risks Associated with Smoking Cigarettes with Low Machine-Measured Yields of Tar and Nicotine". In:

Sinharay, Sandip and Hal S. Stern (2005). "An empirical comparison of methods for computing Bayes factors in generalized linear mixed models". In: *Journal of Computational and Graphical Statistics* 14.2, pp. 415–435. DOI: 10.1198/106186005X47471.

Smith, A.F.M., A.M. Skene, J.E.H. Shaw, J.C. Naylor, and M. Dransfield (1985). "The implementation of the Bayesian paradigm". In: *Communications in Statistics: Theory and Methods* 14.5, pp. 1079–1102. DOI: 10.1080/03610928508828963.

Spiegelhalter, David J., Nicola G. Best, Bradley P. Carlin, and Angelika van der Linde (2002). "Bayesian measures of model complexity and fit". In: *Journal of the Royal Statistical Society: Series B* 24.4, pp. 583–616. DOI: 10.1111/1467-9868.00353.

Spiegelhalter, David, Andrew Thomas, Nicky Best, and Dave Lunn (2014). *OpenBUGS User Manual*. Version 3.2.3. URL: http://www.openbugs.net/Manuals/Manual.html (visited on).

Su, Chun-Lung and Wesley O. Johnson (2006). "Large-sample joint posterior approximations when full conditionals are approximately normal: application to generalized linear mixed models". In: *Journal of the American Statistical Association* 101.474, pp. 795–811. DOI: 10.1198/016214505000001311.

Suwan-Ampai, Plernpit, Ana Navas-Acien, Paul T. Strickland, and Jacqueline Agnew (2009). "Involuntary tobacco smoke exposure and urinary levels of polycyclic aromatic hydrocarbons in the United States, 1999 to 2002". In: *Cancer Epidemiology, Biomarkers and Prevention* 18.3, pp. 884–893. DOI: 10.1158/1055-9965.EPI-08-0939.

Thurmond, M.C., A.J. Branscum, W.O. Johnson, E.J. Bedrick, and T.E. Hanson (2005). "Predicting the probability of abortion in dairy cows: a hierarchical Bayesian logistic-survival model using sequential pregnancy data". In: *Preventive Veterinary Medicine* 68.2, pp. 223–239. DOI: 10.1016/j.prevetmed.2005.01.008.

Watanabe, Sumio (2010a). "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory". In: *Journal of machine learning research* 11, pp. 3571–3594.

— (2010b). "Equations of states in singular statistical estimation". In: *Neural Networks* 23.1, pp. 20–34. DOI: 10.1016/j.neunet.2009.08.002.

Williams, D.A. (1982). "Extra-binomial variation in logistic linear models". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 31.2, pp. 144–148.

Windham, Gayle C., E. P. Elkin, S.H. Swan, K.O. Waller, and L. Fenster (1999). "Cigarette smoking and effects on menstrual function". In: *Obstetrics and Gynecology* 93.1, pp. 59–65. DOI: 10.1016/S0029-7844(98)00317-2.

Windham, Gayle C., Patrick Mitchell, Meredith Anderson, and Bill L. Lasley (2005). "Cigarette Smoking and Effects on Hormone Function in Premenopausal Women". In: *Environmental Health Perspectives* 113.10, pp. 1285–1290. DOI: 10.1289/ehp.7899.

Wolfe, Phillip (1969). "Convergence conditions for ascent methods". In: *SIAM Review* 11.2, pp. 226–235. DOI: 10.1137/1011036.

Zhang, Guoyi, Fletcher Christensen, and Wei Zheng (2015). "Nonparametric regression estimators in complex surveys". In: *Journal of Statistical Computation and Simulation* 85.5, pp. 1026–1034. DOI: 10.1080/00949655.2013.860139.

Zumoff, Barnett, Lorraine Miller, Charles D. Levit, Ellen H. Miller, et al. (1990). "The effect of smoking on serum progesterone, estradiol, and luteinizing hormone levels over a menstrual cycle in normal women". In: *Steroids* 55.11, pp. 507–511. DOI: 10.1016/0039-128X(90)90089-T.

Appendix A

Appendix

A.1 Complete the Square

Proposition 3.1. *For conformable vectors X, μ_1, and μ_2; and for conformable symmetric matrices A_1 and A_2;*

$$\begin{aligned}&(X-\mu_1)^T A_1(X-\mu_1)+(X-\mu_2)^T A_2(X-\mu_2)\\&\quad=(X-\mu^*)^T(A_1+A_2)(X-\mu^*)+(\mu_1-\mu_2)^T A_1(A_1+A_2)^{-1}A_2(\mu_1-\mu_2),\end{aligned}$$

where $\mu^*=(A_1+A_2)^{-1}(A_1\mu_1+A_2\mu_2)$.

Proof. Let $\mu^* = (A_1 + A_2)^{-1}(A_1\mu_1 + A_2\mu_2)$. Then

$$\begin{aligned}
&(X-\mu_1)^T A_1(X-\mu_1) + (X-\mu_2)^T A_2(X-\mu_2) \\
&\quad = X^T(A_1+A_2)X + \mu_1^T A_1\mu_1 + \mu_2^T A_2\mu_2 - X^T(A_1\mu_1 + A_2\mu_2) - (A_1\mu_1 + A_2\mu_2)^T X \\
&\qquad + \mu^{*T}(A_1+A_2)X - \mu^{*T}(A_1+A_2)X \\
&\qquad + X^T(A_1+A_2)\mu^* - X^T(A_1+A_2)\mu^* \\
&\qquad + \mu^{*T}(A_1+A_2)\mu^* - \mu^{*T}(A_1+A_2)\mu^*.
\end{aligned}$$

Now observe that

$$\begin{aligned}
&\mu^{*T}(A_1+A_2)X + +X^T(A_1+A_2)\mu^* \\
&\quad = (A_1\mu_1 + A_2\mu_2)^T(A_1+A_2)^{-1}(A_1+A_2)X + X^T(A_1+A_2)(A_1+A_2)^{-1}(A_1\mu_1 + A_2\mu_2) \\
&\quad = (A_1\mu_1 + A_2\mu_2)^T X + X^T(A_1\mu_1 + A_2\mu_2).
\end{aligned}$$

This allows us to simplify the original equation to

$$\begin{aligned}
&(X-\mu_1)^T A_1(X-\mu_1) + (X-\mu_2)^T A_2(X-\mu_2) \\
&\quad = (X-\mu^*)^T(A_1+A_2)(X-\mu^*) + \mu_1^T A_1\mu_1 + \mu_2^T A_2\mu_2 - \mu^{*T}(A_1+A_2)\mu^*.
\end{aligned}$$

It is sufficient, then, to show that

$$\mu_1^T A_1\mu_1 + \mu_2^T A_2\mu_2 - \mu^{*T}(A_1+A_2)\mu^* = (\mu_1-\mu_2)^T A_1(A_1+A_2)^{-1}A_2(\mu_1-\mu_2).$$

First, observe that

$$\mu^{*T}(A_1+A_2)\mu^* = \mu^{*T}(A_1\mu^* + A_2\mu^*) = \mu^{*T}A_1\mu^* + \mu^{*T}A_2\mu^*.$$

Then we can write

$$\begin{aligned}
&\mu_1^T A_1 \mu_1 + \mu_2^T A_2 \mu_2 - \mu^{*T}(A_1 + A_2)\mu^* \\
&\quad = (\mu_1 - \mu^*)^T A_1 (\mu_1 - \mu^*) + (\mu_2 - \mu^*)^T A_2 (\mu_2 - \mu^*).
\end{aligned}$$

But

$$\begin{aligned}
\mu_1 - \mu^* &= (A_1 + A_2)^{-1}(A_1\mu_1 + A_2\mu_1) - (A_1 + A_2)^{-1}(A_1\mu_1 + A_2\mu_2) \\
&= (A_1 + A_2)^{-1} A_2 (\mu_1 - \mu_2), \\
\mu_2 - \mu^* &= (A_1 + A_2)^{-1}(A_1\mu_2 + A_2\mu_2) - (A_1 + A_2)^{-1}(A_1\mu_1 + A_2\mu_2) \\
&= (A_1 + A_2)^{-1} A_1 (\mu_2 - \mu_1).
\end{aligned}$$

This allows us to rewrite as follows.

$$\begin{aligned}
&\mu_1^T A_1 \mu_1 + \mu_2^T A_2 \mu_2 - \mu^{*T}(A_1 + A_2)\mu^* \\
&\quad = (\mu_1 - \mu_2)^T A_2 (A_1 + A_2)^{-1} A_1 (A_1 + A_2)^{-1} A_2 (\mu_1 - \mu_2) \\
&\qquad + (\mu_1 - \mu_2)^T A_1 (A_1 + A_2)^{-1} A_2 (A_1 + A_2)^{-1} A_1 (\mu_1 - \mu_2) \\
&\quad = (\mu_1 - \mu_2)^T \left[A_2 (A_1 + A_2)^{-1} A_1 (A_1 + A_2)^{-1} A_2 + A_1 (A_1 + A_2)^{-1} A_2 (A_1 + A_2)^{-1} A_1 \right] (\mu_1 - \mu_2).
\end{aligned}$$

But

$$
\begin{aligned}
&A_2(A_1+A_2)^{-1}A_1(A_1+A_2)^{-1}A_2 + A_1(A_1+A_2)^{-1}A_2(A_1+A_2)^{-1}A_1 \\
&= A_1(A_1+A_2)^{-1}A_2(A_1+A_2)^{-1}A_1 + A_2(A_1+A_2)^{-1}A_1(A_1+A_2)^{-1}A_2 \\
&\qquad + A_1(A_1+A_2)^{-1}A_2(A_1+A_2)^{-1}A_2 - A_1(A_1+A_2)^{-1}A_2(A_1+A_2)^{-1}A_2 \\
&= A_1(A_1+A_2)^{-1}A_2(A_1+A_2)^{-1}(A_1+A_2) \\
&\qquad + A_2(A_1+A_2)^{-1}A_1(A_1+A_2)^{-1}A_2 - A_1(A_1+A_2)^{-1}A_2(A_1+A_2)^{-1}A_2 \\
&= A_1(A_1+A_2)^{-1}A_2 + \left(A_2(A_1+A_2)^{-1}A_1 - A_1(A_1+A_2)^{-1}A_2\right)(A_1+A_2)^{-1}A_2 \\
&= A_1(A_1+A_2)^{-1}A_2 \\
&\qquad + \left(A_2(A_1+A_2)^{-1}A_1 - A_1(A_1+A_2)^{-1}A_2\right)(A_1+A_2)^{-1}A_2 \\
&\qquad + \left(A_1(A_1+A_2)^{-1}A_1 - A_1(A_1+A_2)^{-1}A_1\right)(A_1+A_2)^{-1}A_2 \\
&= A_1(A_1+A_2)^{-1}A_2 + \left((A_1+A_2)(A_1+A_2)^{-1}A_1 - A_1(A_1+A_2)^{-1}(A_1+A_2)\right)(A_1+A_2)^{-1}A_2 \\
&= A_1(A_1+A_2)^{-1}A_2 + (A_1 - A_1)(A_1+A_2)^{-1}A_2 \\
&= A_1(A_1+A_2)^{-1}A_2.
\end{aligned}
$$

Therefore

$$\mu_1^T A_1 \mu_1 + \mu_2^T A_2 \mu_2 - \mu^{*T}(A_1+A_2)\mu^* = (\mu_1-\mu_2)^T A_1(A_1+A_2)^{-1}A_2(\mu_1-\mu_2).$$

□

A.2 Gauss-Hermite Quadrature

Let $\mathbf{Y} = \{Y_i\} = \{y_{ij}\}$ be a $kn \times 1$ vector of response data on clusters $i \in \{1, ..., k\}$, with $j \in \{1, ..., n\}$ observations per cluster. We use a balanced design with common n for all clusters to simplify some of the following linear algebra, but the results we obtain do not require this balance.

Let β be a $p \times 1$ vector of regression parameters. Let $\mathbf{X}$ be the $kn \times p$ design matrix for the regression parameters, $\mathbf{X}_i$ be the $n \times p$ block of the $\mathbf{X}$ matrix corresponding to cluster i, and X_{ij} be the $1 \times p$ row vector corresponding to the j^{th} observation on cluster i.

Let $\boldsymbol{\gamma} = \begin{bmatrix} \gamma_1^T & ... & \gamma_k^T \end{bmatrix}^T$ be a $kq \times 1$ vector of random effects, with γ_i the $q \times 1$ vector of random effects corresponding to cluster i. Let $\mathbf{Z}$ be the $kn \times kq$ block diagonal design matrix for the random effects. Let $\mathbf{Z}_i$ be the $n \times q$ submatrix of $\mathbf{Z}$ corresponding to its i^{th} diagonal block, and Z_{ij} be the $1 \times q$ row vector corresponding to the j^{th} row of the $\mathbf{Z}_i$ matrix.

Let $\boldsymbol{\psi} = \begin{bmatrix} \psi_1^T & ... & \psi_k^T \end{bmatrix}^T$ be the mean of the random effects vector $\boldsymbol{\gamma}$, and let $\boldsymbol{\Sigma}$ be block diagonal Σ_i, $i \in \{1, ..., k\}$ be the covariance matrix of the random effects. We assume that $\boldsymbol{\gamma} \sim \mathrm{N}_{kq}(\boldsymbol{\psi}, \boldsymbol{\Sigma})$, or equivalently here that $\gamma_i \overset{\text{indep}}{\sim} \mathrm{N}_q(\psi_i, \Sigma_i)$. We use θ to refer to the collection of parameters $\{\beta, \boldsymbol{\psi}, \boldsymbol{\Sigma}\}$.

Gauss-Hermite quadrature requires transforming γ_i to a multivariate standard normal distribution. Define, therefore,

$$\nu_i = \Sigma_i^{-1/2}(\gamma_i - \psi_i), \qquad \gamma_i = \Sigma_i^{1/2}\nu_i + \psi_i.$$

Then the joint density of Y_i and ν_i is

$$f_{\nu_i}(Y_i, \nu_i \mid \theta) = f_{\gamma_i}\left(Y_i, (\Sigma_i^{1/2}\nu_i + \psi_i) \mid \theta\right) |\Sigma_i|^{-1/2}$$

where ν_i is also a $q \times 1$ vector, like γ_i. We will further write

$$\nu_i = \begin{bmatrix} \nu_{i1} \\ \vdots \\ \nu_{iq} \end{bmatrix}$$

Let $u_H^{(s)}$, $s \in \{1, ..., B_n\}$ be the roots of the n^{th} degree Hermite polynomial—the nodes on which the function will be evaluated. Let $v_H^{(s)}$, $s \in \{1, ..., B_n\}$ be their associated weights. Then the Gauss-Hermite quadrature approximation to $f(\mathbf{Y} \mid \theta)$ is given by

$$
\begin{aligned}
f(\mathbf{Y} \mid \theta) &= \prod_{i=1}^{k} f(Y_i \mid \theta) \\
&= \prod_{i=1}^{k} \int f_{\nu_i}(Y_i, \nu_i \mid \theta) d\nu_i \\
&= \prod_{i=1}^{k} \int f_{\gamma_i}\left(Y_i, (\Sigma_i^{1/2}\nu_i + \psi_i) \mid \theta\right) d\nu_i \, |\Sigma_i|^{-1/2} \\
&\doteq \prod_{i=1}^{k} \left(\sum_{s_1=1}^{B_n} \cdots \sum_{s_q=1}^{B_n} f\left(Y_i, \left(\Sigma_i^{1/2} \begin{bmatrix} u_H^{(s_1)} \\ \vdots \\ u_H^{(s_q)} \end{bmatrix} + \psi_i \right) \mid \theta \right) |\Sigma_i|^{-1/2} \, v_H^{(s_1)} \ldots v_H^{(s_q)} \right)
\end{aligned}
$$

These approximations to the marginal density can then be applied to a posterior sample $\theta^{(s)}$, $s \in \{1, ..., B_t\}$ and used with the equations

$$
\begin{aligned}
p_{Dm} &\doteq -\frac{2}{B} \sum_{s=1}^{B} \log f\left(\mathbf{Y} \mid \theta^{(s)}\right) + 2 \log f\left(\mathbf{Y} \mid \hat{\theta}\right) \\
DIC_m &\doteq -\frac{4}{B} \sum_{s=1}^{B} \log f\left(\mathbf{Y} \mid \theta^{(s)}\right) + 2 \log f\left(\mathbf{Y} \mid \hat{\theta}\right)
\end{aligned}
$$

to obtain numerical approximations for p_{Dm} and DIC_m.

A.3 Taylor Approximation Error in REO GLMMs

Of interest when approximating a quantity is the degree to which the approximation diverges from the actual quantity. As we see in Equation (4.8),

$$f(Y_i, \gamma_i \mid \theta) = \hat{f}(Y_i, \gamma_i \mid \theta) \exp\left(r(g, \gamma_i, \hat{\gamma}_i)\right),$$

where $r(g, \gamma_i, \hat{\gamma}_i) = o(||\gamma_i - \hat{\gamma}_i||^2)$ as $n \to \infty$. If the log-likelihood $\ell\left(\gamma_i, \theta \mid Y_i\right)$ is thrice-differentiable in γ_i, as is true for all common exponential family densities, then we can further express the remainder in the Lagrange mean-value form, Equation (4.6),

$$r(g, \gamma_i, \hat{\gamma}_i) = \frac{g^{(3)}(\xi_L)}{3!} \left(\gamma_i - \hat{\gamma}_i\right)^3,$$

where ξ_L is some real number between γ_i and $\hat{\gamma}_i$.

In large samples, the behavior of the remainder is governed by Proposition 4.1. Under mild conditions—for each cluster i, the elements of Y_i must be conditionally independent and have an expected value, and $\ddot{g}(t)$ must be continuous on a closed neighborhood near γ_i—$\hat{f}(Y_i, \gamma_i \mid \theta) \xrightarrow{\text{a.s.}} f(Y_i, \gamma_i \mid \theta)$. Further, we know $\hat{f}(\gamma_i \mid Y_i, \theta)$ is a Normal density. Unfortunately we have no guarantee that $\hat{f}(Y_i, \gamma_i \mid \theta)$ and $\hat{f}(Y_i \mid \theta)$ are proper densities, or even that they correspond to finite measures, except at the limit. Proposition 4.2 below gives conditions sufficient for us to know that the approximate marginal density corresponds to a finite measure. Corollary 4.3 establishes that $\hat{f}(Y_i \mid \theta)$ corresponds to a finite measure iff $\hat{f}(Y_i, \gamma_i \mid \theta)$ does as well; and that they share the same constant.

Proposition A.1.

$$\hat{F}(c \mid \theta) = \int_{-\infty}^{c} \hat{f}(t \mid \theta) dt$$

is a finite measure if real-valued functions $M_1(\theta)$ and $M_2(\theta)$ exist such that

$$0 \leq M_1(\theta) \leq \mathrm{E}_{\gamma_i|Y_i,\theta}[\exp\left(r(g,\gamma_i,\hat{\gamma}_i(Y_i))\right)] \leq M_2(\theta).$$

Proof. Observe that

$$f(Y_i,\gamma_i \mid \theta) = \hat{f}(\gamma_i \mid Y_i,\theta)\hat{f}(Y_i \mid \theta)\exp\left(r(g,\gamma_i,\hat{\gamma}_i(Y_i)\right).$$

Then

$$\begin{aligned}
1 &= \int\int f(t,s \mid \theta)dsdt \\
&= \int\int \hat{f}(s \mid t,\theta)\hat{f}(t \mid \theta)\exp\left(r(g,s,\hat{\gamma}_i(t)\right)dsdt \\
&= \int \hat{f}(t \mid \theta)\left(\int \hat{f}(s \mid t,\theta)\exp\left(r(g,s,\hat{\gamma}_i(t)\right)ds\right)dt \\
&= \int \hat{f}(t \mid \theta)\,\mathrm{E}_{\gamma_i|Y_i,\theta}[\exp\left(r(g,\gamma_i,\hat{\gamma}_i(t))\right)]\,dt,
\end{aligned}$$

since $\hat{f}(\gamma_i \mid Y_i,\theta)$ is a proper Normal density.

Then since $M_1(\theta) \leq \mathrm{E}_{\gamma|Y_i,\theta}[r(g,\gamma_i,\hat{\gamma}_i(Y_i)] \leq M_2(\theta)$ and $\hat{f}(Y_i \mid \theta)$ is strictly positive,

$$M_1(\theta)\int \hat{f}(Y_i \mid \theta)dY_i \leq \int \hat{f}(Y_i \mid \theta)\,\mathrm{E}_{\gamma|Y_i,\theta}[r(g,\gamma_i,\hat{\gamma}_i(Y_i)]\,dY_i \leq M_2(\theta)\int \hat{f}(Y_i \mid \theta)dY_i,$$

and thus

$$\frac{1}{M_2(\theta)} \leq \int \hat{f}(Y_i \mid \theta) \leq \frac{1}{M_1(\theta)}$$

since the middle integral evaluates to 1. □

Corrolary A.2. *$\hat{F}(Y_i \mid \theta)$ is a finite measure iff*

$$\hat{F}(Y_i,\gamma_i \mid \theta) = \int_{-\infty}^{Y_i}\int_{-\infty}^{\gamma_i} \hat{f}(t,s \mid \theta)dsdt$$

is also a finite measure. Further, $\hat{F}(Y_i \mid \theta)$ and $\hat{F}(Y_i, \gamma_i \mid \theta)$ share the same normalizing constant.

Proof. Set $w = \int \hat{f}(t \mid \theta)dt$. Then, since $\hat{f}(\gamma_i \mid Y_i, \theta)$ is a Normal density, we have

$$\begin{aligned} w &= \int \hat{f}(t \mid \theta) \left(\int \hat{f}(s \mid t, \theta)ds \right) dt \\ &= \int \int \hat{f}(t \mid \theta)\hat{f}(s \mid t, \theta)dsdt \\ &= \int \int \hat{f}(t, s \mid \theta)dsdt. \end{aligned}$$

□

Simulations have shown that $\hat{f}(Y_i \mid \theta)$ and $\hat{f}(Y_i, \gamma_i \mid \theta)$ are well behaved relative to the true joint and marginal densities (see Section 4.4.2). We have not, however, been able to show that the condition for Proposition A.1 holds in general for exponential family distributions or specifically for the binomial or Poisson. This is an element of future work we intend to pursue. Even if we cannot prove that these functions are guaranteed to have finite measure, the results we obtain in simulations show that our approximation is reasonably accurate. Further, that Edgeworth expansions are not guaranteed to yield probability measures. Although we would prefer to know that our approximations correspond to finite measures, we do not consider our inability to achieve this result especially limiting.