ABSTRACT

Title of Dissertation:	ADJUSTMENT FOR DENSITY METHOD TO ESTIMATE RANDOM EFFECTS IN HIERARCHICAL BAYES MODELS
	Lijuan Cao, Doctor of Philosophy, 2018
Dissertation directed by:	Professor, Partha Lahiri, Department of Mathematics & Joint Program in Survey Methodology

The Adjustment for Density Method (ADM) has received considerable attention in recent years. The method was proposed about thirty years back in approximating a complex univariate density by a density from the Pearson family of distributions. The ADM has been developed to approximate posterior distributions of hyper-parameters, shrinkage parameters and random effects of a few well-known univariate hierarchical Bayesian models. This dissertation advances the ADM to approximate posterior distributions of hyper-parameters, shrinkage parameters, synthetic probabilities and multinomial probabilities associated with a multinomial-Dirichlet-logit Bayesian hierarchical model. The method is adapted so it can be applied to weighted counts. We carefully propose prior for the hyper-parameters of the multinomial-Dirichlet-logit model so as to ensure propriety of posterior of relevant parameters of the model and to achieve good small sample properties. Following general guidelines of the ADM for univariate distributions, we devise suitable adjustments to the posterior density of the

hyper-parameters so that adjusted posterior modes lie in the interior of the parameter space and to reduce the bias in the point estimates. Beta distribution approximations are employed when approximating the posterior distributions of the individual shrinkage factors and Dirichlet distribution approximations are used when approximating the posterior distributions of the synthetic probabilities. The parameters of the beta or the Dirichlet posterior density are approximated carefully so the method approximates the exact posterior densities accurately. Simulation studies demonstrate that our proposed approach in estimating the multinomial probabilities in the multinomial-Dirichlet-logit model is accurate in estimation, fast in speed and has better operating characteristics compared to other existing procedures. We consider two applications of our proposed hierarchical Bayes model using complex survey and Big Data. In the first example, we consider small area gender proportions using a binomial-beta-logit model. The proposed method improves on a rival method in terms of smaller margins of error. In the second application, we demonstrate how small area multi-category race proportions estimates, obtained by direct method applied on Twitter data, can be improved by the proposed method. This dissertation ends with a discussion on future research in the area of ADM.

ADJUSTMENT FOR DENSITY METHOD TO ESTIMATE RANDOM EFFECTS IN HIERARCHICAL BAYES MODELS

by

Lijuan Cao

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Professor Partha Lahiri, Chair
Professor Eric Slud
Professor Paul Smith
Professor Laura Stapleton
Assistant Professor Kunpeng Zhang

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor Professor Partha Lahiri for the continuous support of my Ph.D. study and research, for his patience, motivation, and immense knowledge. He introduced me to the interesting research area in this dissertation and encouraged me to accumulate knowledge and explore in this area. His guidance helped me in all the time of research and writing of this thesis. I could not have finished the research project and writing this thesis without his help.

Besides my advisor, I would like to thank the rest of my thesis committee: Dr. Eric Slud, Dr. Paul Smith, Dr. Laura Stapleton and Dr. Kunpeng Zhang. The final version of my thesis has improved greatly because of their patient proofreading, insightful comments and intellectually rigorous critique. I want to give special thanks to Dr. Eric Slud and Dr. Paul Smith for their valuable teachings. I have built up a solid foundation for scientific researches in statistics.

I shall also thank my fellow colleague and officemate Wenbo Li for proofreading the mathematical proofs in this thesis and the help I have received from him throughout the program. I do not hesitate to express thanks to my dear friends outside the PhD program Xuejing Wang, Yibo Gu and Alice Hu for their

encouragement and support in some difficult times. Sincere thanks will also be given to the previous and current graduate coordinators Celeste Regalado and Cristina Garcia for their excellent work and many answers to my questions about the administrative related issues.

Last but not the least, I would like to thank my family: my parents and my cousin Liyan Tang for supporting me spiritually throughout writing this thesis and my life in general.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

This dissertation advances the Adjustment for Density Method (ADM) in approximating the posterior distributions of hyper-parameters, shrinkage parameters, synthetic probabilities and multinomial probabilities associated with a multinomial-Dirichlet-logit model and demonstrates its advantages over some existing methods through real data examples and simulation studies. The multinomial-Dirichlet-logit model is proposed to combine information from multiple data sources. The ADM based on this model produces estimates of the multinomial probabilities that are much more reliable than the direct multinomial sample proportions for small areas. The individual shrinkage factors allow the small area proportion estimates to shrink away from the direct proportion estimates towards the synthetic proportion estimates obtained by the multinomial logistic regression. We carefully propose a hyper-prior for the hyper-parameters of the multinomial-Dirichlet-logit model so as to ensure propriety of posterior of relevant parameters of the model and to achieve good small sample properties. Following general guidelines of the ADM for univariate distributions, we devise suitable adjustments to the posterior density of the hyper-parameters so that posterior modes lie in the interior of the parameter space and to reduce the bias in the point estimates. Beta distribution

approximations are employed when approximating the posterior distributions of the individual shrinkage factors and Dirichlet distribution approximations are used when approximating the posterior distributions of the synthetic probabilities. Using Monte Carlo simulations and data analysis, we demonstrate that the proposed ADM yields good point estimates, variance estimates and approximate distributions for all parameters of interest. Compared with the MCMC, the available program for the ADM is fast enough to be used interactively for model checking purpose. The ADM introduces a third level hyper-prior on the hyper-parameters to prevent infinite value for the variance component estimate, which the MLE of the variance component occasionally takes on. Through the restricted maximum likelihood (REML) type correction to the posterior distribution function of the hyper-parameters, the bias in the variance component estimate is corrected. The ADM approximations are applied when approximating the posterior distributions of the shrinkage factors by beta distributions and the synthetic proportions by Dirichlet distributions. The ADM generates closed-forms for the posterior means and the posterior variances for the multinomial probabilities, with the variances in the hyper-parameter estimates incorporated. The resulting wider interval estimates for the multinomial probabilities partly explain the higher interval coverage rates. The ADM improves on the operating characteristics (e.g., risk and coverage rate) of the multinomial probability estimates compared with the EB plug-in methods.

Empirical Bayes and hierarchical Bayes models are useful in data analysis, thus there exists an incentive to improve the parameter estimation procedure for these models. This dissertation gives two data analysis examples using hierarchical Bayes models. In one of the examples, we introduce the Twitter direct gender proportion estimates as the auxiliary variable to the American Community Survey (ACS) gender counts in small areas in a hierarchical binomial regression model. This example proposes an alternative method to calculate the point estimates and variance estimates of the small area gender proportions. The alternative gender proportion estimates have smaller margins of error than the direct small area estimates. In the other example, we apply the proposed ADM in estimating the posterior means and posterior variances of the multinomial probabilities in a multinomial-Dirichlet-logit model as mentioned above to a Twitter race count dataset. There are some small areas in the Twitter data with small race counts and the direct proportion estimates for these areas cannot be trusted. The proposed procedure generates synthetic proportion estimates by multinomial logistic regression on area-wise predictors and permits the proportion estimates to shrink between the direct estimates and the synthetic estimates. The extent of the shrinkages depend on the small area sample sizes.

There are huge literature in small area estimation[42]. in empirical and hierarchical Bayes models. Ghosh and Rao (1994) reviewed the multi-level Bayes models to estimate county population and small area per capita income (PCI)

and to adjust for population undercount in the 1980 U.S. Census[17]. The estimates in the examples of the small area PCI and the adjustment for population undercount are weighted averages of the sample estimates and the synthetic regression estimates. By introducing the synthetic regression estimates, the updated estimates borrow strength from related areas. More recently, Rao and Molina (2015) discussed the issues in empirical and hierarchical Bayes models in small area estimation using the MCMC approach[42].

There are papers on Bayesian and empirical Bayesian methods for multinomial-Dirichlet models. Carlin and Louis (2010) briefly discussed the MCMC estimation in multinomial-Dirichlet Bayes models[9]. In order to obtain reliable estimates from American Community Survey (ACS) to determine whether to provide language assistance during elections for designated language-minority groups of citizens who are unable to speak or understand English well enough to participate in the electoral process, Ashmead and Slud (2017) applied a multinomial-Dirichlet model to carry out inference on the small area proportions of four categories of voting-age citizens and proposed model selection and model validation procedures[1]. Slud and Ashmead (2017) also developed a hybrid method to estimate the variances of the proportion estimates in a multinomial-Dirichlet hierarchical model[47]. Multinomial-Dirichlet-logit model can also be applied to estimate small area multi-category proportions (e.g., race and employment status).

Because of the demand of analysis tools for multinomial data, we improve the

parameter estimation for a multinomial-Dirichlet-logit model. The parameter estimation method is extended from Morris and his team's research in ADM. The ADM generates estimates of random effects with good quality even when the sample size is small with fast computation speed. However, they restrict themselves only to the case of univariate distributions including normal, binomial and Poisson hierarchical models and they have not conducted research on ADM for multivariate distribution such as the multinomial-Dirichlet-logit model we consider in this dissertation. The ADM requires case by case detailed extension to papers written by Morris and his collaborators to a new model, including the selection of suitable hyper-prior, proof of conditions for posterior propriety and selection of the suitable approximating distributions to exact posterior distributions. All of the model-specific technical work involved in ADM make this multinomial problem a non-trivial problem.

There are some existing methods to implement Bayesian methodology for multi-level models. One of the most commonly used methods is MCMC. This method requires checking the convergence of the MCMC and is computer intensive. In case of big data, this method is extremely time-consuming. Since MCMC is a stochastic procedure, estimates vary even when the same model is applied to the same dataset repeatedly. This randomness is not favored by legal and public policy applications[13;38]. Moreover, MCMC is not convenient in handling non-integer weighted counts as commonly observed in survey data since some existing MCMC packages require integer counts

for multinomial distribution. Another parameter estimation procedure is the empirical Bayes (EB) procedure with the MLEs or the restricted maximum likelihood (REML) estimates of the hyper-parameters plugged in. The two EB procedures are denoted by EB-MLE and EB-REML, respectively. In the multinomial-Dirichlet-logit model in this dissertation, the MLE and the REML estimate of the variance component will occasionally occur at an infinite value for small sample sizes. This is consistent with research obtained by Christiansen and Morris (1997)[13]. The infinite variance component estimate will cause trouble in the inference of the multinomial probabilities. Also, these two EB-plugin procedures are not ideal in the sense of low coverage rates for small areas and coverage rates varying with area sample sizes. All these disadvantages of the EB-MLE and the EB-REML methods have been observed in the simulation studies in Chapter 4.

To address all the problems in the MCMC and EB procedures, we propose the ADM to improve the parameter estimation in the multinomial-Dirichlet-logit model. The ADM serving as an alternative parameter estimation procedure to MCMC was introduced by Morris (1988)[37]. Ever since, Morris and his students have written a series of papers in this field. They have developed the ADM to estimate the first level random effects in multi-level Poisson[13], normal and skewed-normal[24;38] and binomial[25] models. Morris and Tang (2011)[38] summarize seven advantages of the ADM, some of which have become obsolete (e.g., (4) in the conclusion is not true any more since MCMC can handle multi-

level generalized linear models at the time of writing this dissertation). We would like to summarize some advantages of ADM observed from our research, including (1) the overwhelmingly fast speed of ADM compared with MCMC permits it to be used repeatedly for model selection, model checking and operating characteristics checking; (2) same results each time the procedure is applied to the same dataset using the same software platform on the same machine while it is impossible for stochastic approximations[13;38]; (3) preventing infinite values for the variance component estimate by adjusting the likelihood of the hyper-parameters; (4) introducing the variance of the hyper-parameter estimates to the random effect estimates and consequently increasing the coverage rates to nearly the nominal coverage rates; (5) the closed-form approximations to the posterior distributions of all the parameters in the hierarchical models; and (6) capability to handle weighted non-integer counts in Poisson, binomial and multinomial models although this advantage is not emphasized in the ADM since some hierarchical Bayes and empirical Bayes methods are applied to non-integer survey-weighted estimates of integer counts[22;32].

This dissertation is structured as follows. Chapter 2 is a review of the ADM papers and dissertations and the readers can skip this chapter if desired. This chapter may be useful to those who may not be familiar with ADM as the research has mostly been conducted by Morris and his students. Chapter 3 and Chapter 4 are the two most important chapters of this dissertation. Chapter 3 details the development of the ADM for a multinomial-Dirichlet-logit hierar-

chical model. Chapter 4 is the study of the proposed ADM through real data examples and simulation experiments. Chapter 5 contains two application examples using hierarchical Bayes models implemented by the ADM. Chapter 6 concludes the dissertation and briefly introduces future research. All the proofs are deferred to the appendix.

Chapter 2 begins with an introduction to the theory of ADM proposed by Morris[37]. The ADM approximates the distribution of the parameter of interest by one of the Pearson family distributions. The selection of the Pearson family distribution is guided the support of the distribution to be approximated. When there are multiple Pearson distributions that satisfy this criterion, the selection of the Pearson distribution can be made based on the performance of the ADM for a particular inferential problem (e.g., the coverage rate as discussed in Chapter 4 of this dissertation). The ADM approximation uses the first two derivatives of the logarithm of adjusted posterior density function of the parameter of interest. This is analogous to the normal approximation with moment matching. Then Chapter 2 details the development of ADM for multi-level Poisson, normal and skewed-normal and binomial models and ends with a discussion of the advantages of the ADM in parameter estimation for hierarchical Bayes models.

Chapter 3 develops the ADM for the multinomial-Dirichlet-logit model. As mentioned earlier, the ADM uses a series of adjustments to the posterior distribution of the hyper-parameters, the shrinkage factors and the proportions.

This chapter first describes the multinomial-Dirichlet-logit model in two mathematically equivalent forms - the descriptive form and the inferential form - both of which have a third level hyper-prior on the hyper-parameters. Then we provide a mild sufficient condition of the data for the posterior distribution of the hyper-parameters to be proper. Once the mild condition of the data is satisfied, the hyper-parameter estimates lie in the interior of the parameter space. Then we use normal approximation to assign a joint normal distribution to the hyper-parameters. Our proposed approximations to the posterior means and posterior variances of the multinomial probabilities take into account the variabilities of the hyper-parameters. Both the posterior means and the posterior variances are functions of the posterior moments of the shrinkage factors and the synthetic probabilities. The distributions of the shrinkage factors are approximated by the beta distributions and the distributions of the synthetic probabilities are approximated by the Dirichlet distributions using ADM. In this chapter we extend the ADM to a multinomial-Dirichlet-logit model. The detailed research includes the selection of the appropriate hyper-prior, the proof of posterior propriety conditions for such a hyper-prior, the approximation of the determinant of the Hessian matrix and the Dirichlet approximations to the posterior distributions of the synthetic proportions.

Chapter 4 compares our proposed ADM with the corresponding hierarchical Bayes method implemented through MCMC and the EB methods. The computational speed of our proposed ADM is overwhelmingly faster than that of

MCMC. For a dataset with 10 areas and 5 categories for each of 2 covariates, the ADM is hundreds of times faster than the MCMC method. The comparison with the EB methods by simulation studies demonstrates that our procedure ensures parameter estimates to lie in the interior of parameter space under a mild condition on the data and has better risks and better coverage rates.

Chapter 5 gives two application examples using hierarchical Bayes models. The first example introduces the Twitter direct small area estimates of the gender proportions as an auxiliary variable to the ACS small area gender counts in a binomial-beta-logit regression model. The small area gender proportion estimates are generated by the ADM developed by Tak, Kelly and Morris (2016)[25] and the estimation procedure is implemented using the Rgbp package[25] in R. It has been verified that our proposed small area gender proportion estimates have smaller margins of error than the ACS estimates. The second application example is to apply our proposed ADM for the multinomial-Dirichlet-logit regression model to the Twitter small area race count dataset and calculate the small area race proportion estimates. The estimated race proportions are weighted averages of the direct race proportion estimates and the synthetic race proportion estimates.

Chapter 6 concludes this dissertation by summarizing the ADM in parameter estimation in hierarchical Bayes models and its application and discussing our contribution to the area of ADM. Multinomial data is a widely observed data

type in both the public and the private sectors. The multinomial-Dirichlet-logit regression model is useful in estimating small area probabilities. The existing parameter estimation procedures for multinomial-Dirichlet-logit regression model can either be slow in speed (e.g., MCMC) or generate undesirable estimates (e.g., EB). Our proposed ADM overcomes the problems in the existing parameter estimation procedures.

Chapter 2: Literature Review

This chapter is a literature review of the researches which have been conducted in the area of ADM. It is for the readers who are not familiar with ADM. The readers can skip this chapter if you are only interested in the work conducted by the author of this dissertation. This chapter introduces the ADM in random effect estimation in some hierarchical Bayes models and the research which has been conducted mainly by Morris and his students. The ADM for different hierarchical models all contain some adjustments to the posterior distribution of the hyper-parameters and the ADM approximations to some posteriors of the parameters of interest (e.g., the shrinkage factors). The ADM approximation allows approximating the posterior distributions by a range of distributions in the Pearson family. The ADM approximation provides convenience when the density to be approximated is defined on a subset of the set of real numbers. And the ADM have been implemented for several hierarchical Bayes models because of their fast computational speed and good operating characteristics (e.g., risk and coverage rate). This chapter is organized as follows. Section 2.1 introduces the theorem and the advantages of ADM in random effect estimation. Sections 2.2 to 2.4 detail the ADM approximation procedures for random effect estimation in Poisson-gamma, normal-normal, and binomial-

beta hierarchical models. Section 2.5 discusses the possible research topics in both theory and application in this area.

2.1 Adjustment for Density Method

ADM was first proposed by Morris[37] in the year 1988. ADM provides an extension to normal approximation and allows approximation by the Pearson family of distributions[37]. The definition of the Pearson family of distributions is given by Morris in his paper[37].

Definition 2.1 (Pearson Family) *The Pearson family is the natural exponential family (NEF) with the quadratic adjustment factor function* $Q(x) = q_2x^2 + q_1x + q_0 > 0$*, which has density:*

$$f(x) = K_Q(m, \mu_0)e^{-m\int \frac{x-\mu_0}{Q(x)}dx}/Q(x)$$

with $\{x : 0 < Q(x) < \infty\}$*. For fixed adjustment factor function Q(x), this density is a two parameter distribution, denoted by*

$$Pearson(m, \mu_0; Q) = Pearson\left[\mu_0, \frac{Q(\mu_0)}{m - q_2}\right],$$

where μ_0 *and* $\frac{Q(\mu_0)}{m-q_2}$ *are the mean and variance of the Pearson distribution, respectively. The variance* $Var(x) = Q(\mu_0)/(m - q_2)$ *is finite if* $m > q_2$*.*

Table 2.1 lists some distributions in the Pearson family and the relevant pa-

rameters and adjustment factors. As seen from the table, the Pearson family contains distributions lying on various intervals. Usually, we select the approximating Pearson distribution with the same support as the density to be approximated. The first column in Table 2.1 presents the Pearson distribution with traditional parameters which are commonly used by statisticians and the last two columns list the Pearson-type parameters μ_0 and m. And the $Q(x)$ column gives the adjustment factor function.

Table 2.1: Some Pearson Families

Distribution	Density $p(x) \propto$	Range(x)	$Q(x)$	q_2	μ_0	m
$Normal(\mu, \sigma^2)$	$e^{-(x-\mu)^2/2\sigma^2}$	$(-\infty, \infty)$	1	0	μ	σ^{-2}
$Gamma(a, b)$	$x^a e^{-bx}$	$(0, \infty)$	x	0	a/b	b
$Inv-Gamma(a, b)$	$x^{-a-1}e^{-b/x}$	$(0, \infty)$	x^2	1	$a/(b-1)$	$b-1$
$Beta(a, b)$	$x^{a-1}(1-x)^{b-1}$	$(0, 1)$	$x(1-x)$	-1	$a/(a+b)$	$a+b$
$F^*(a, b)$	$\frac{x^a}{(1+x)^{a+b-1}}$	$(0, \infty)$	$x(1+x)$	1	$a/(b-1)$	$b-1$
t_n	$(1+\frac{x^2}{n})^{-\frac{n-1}{2}}$	$(-\infty, \infty)$	$n+x^2$	1	0	$n-1$

Consider approximating the density $f(x)$ by a Pearson family distribution.

$$f(x) \approx K_Q(m, \mu_0) e^{-m \int \frac{x-\mu_0}{Q(x)} dx} / Q(x) \tag{2.1}$$

Then the goal is to estimate the parameters μ_0 and m of the Pearson distribution. By first multiplying both sides of equation (2.1) by $Q(x)$ and then

taking the natural log on both sides, we have

$$\log\Big(f(x)Q(x)\Big) \approx \log\Big(K_Q(m,\mu_0)\Big) - m\int \frac{x-\mu_0}{Q(x)}dx. \tag{2.2}$$

Let $l(x) = \log\Big(f(x)Q(x)\Big)$ and take the first and second derivatives of $l(x)$,

$$\frac{\partial l(x)}{\partial x} = -m\frac{x-\mu_0}{Q(x)} \tag{2.3}$$

$$\frac{\partial^2 l(x)}{\partial x^2} = -\frac{m}{Q(x)} + \frac{m(x-\mu_0)Q'(x)}{Q^2(x)}. \tag{2.4}$$

Set the first derivative in equation (2.3) to be 0 and get the solution $x_0 = \mu_0$. Thus, $x_0 = \mu_0$ maximizes the adjusted density $f(x)Q(x)$, which is the density to be approximated $f(x)$ multiplied by the adjustment factor function $Q(x)$ of the approximating Pearson distribution. Meanwhile, the Pearson distribution parameter μ_0 can be estimated by x_0. That is, use the MLE of the adjusted density to estimate the mean of both the density $f(x)$ and the Pearson distribution and avoid integration. Then by inserting $x_0 = \mu_0$ to the second derivative in equation (2.4), we have $-l''(x_0) = m/Q(x_0)$. This is the Fisher information of the adjusted density $f(x)Q(x)$. And obviously, the second Pearson distribution parameter m can be estimated by $m = -l''(x_0)Q(x_0)$. Thus, the variance of the density $f(x)$ can be estimated by the Pearson distribution variance $Q(\mu_0)/(m - q_2)$. Analogous to normal approximation, the Pearson-type parameters are estimated by the first two derivatives of log adjusted density. That is the reason that this approximation procedure is named ADM. The

normal approximation is actually a special case of ADM with adjustment factor function equal to 1 and the Pearson-type parameters coinciding with the commonly used parameters, that is, $\mu_0 = \mu$ and $m = \sigma^{-2}$. The summary of the procedures for Pearson approximation, also named ADM approximation, are listed in Definition 2.2.

Definition 2.2 (Pearson Approximation) *The Pearson approximation is to approximate a density $f(x)$ by a distribution in the Pearson family. Usually one selects the distribution in the Pearson family which has the same support as $f(x)$.*

The steps of the approximation are:

1. *Let $l(x) = \log\Big(f(x)Q(x)\Big)$, where Q(x) is the adjustment factor function of the approximating Pearson distribution;*

2. *Solve $l'(x) = 0$. The solution x_0 is the MLE of the adjusted density $f(x)Q(x)$ and the estimated mean of both $f(x)$ and the approximating Pearson distribution. Thus, the mean parameter μ_0 of the approximating Pearson distribution can be estimated by x_0;*

3. *The parameter m in the approximating Pearson distribution is equal to the second derivative of $l(x)$ multiplied by the adjustment factor evaluated at x_0: $m = -l''(x_0)Q(x_0)$;*

4. *Solve for the estimates of the traditional parameters of the Pearson distribution based on the relationships between the Pearson-type parameters*

and the traditional parameters. The relationships can be found in the last two columns of Table 2.1.

The ADM approximation has been successful in functioning as an important part of estimating the parameters for a series of hierarchical Bayes models, including Poisson-gamma model[13], normal-normal model[24;38], and binomial-beta model[25]. It has been shown that ADM has multiple advantages over MCMC and other procedures in estimating parameters in hierarchical Bayes models[13;38]. ADM approximates the posterior mean and posterior variance of all the parameters in a hierarchical Bayes model in closed-form. Closed-form allows fast computation and the same estimate each time a model is applied to the same dataset[13;38]. These features will be favored by some data practitioners and cannot be achieved by stochastic approximation procedures such as MCMC[13]. Directly giving closed-forms for the mean and variance estimates avoids the burn-in period in the MCMC method, increasing the computation speed significantly[13]. It is documented that in the Poisson-gamma case the speed is 70 times faster[13] than MCMC and, as reported in Chapter 4 of this dissertation, we have observed a speed which is hundreds of times faster than MCMC for the multinomial-Dirichlet-logit model. Fast computation makes repeated simulations feasible when the operating characteristics of a procedure are of interest. Past work has shown that ADM is superior in generating good operating characteristics (e.g. coverage, interval width and risk functions)[13;24;38] compared with a range of parameter estimation methods. The

rest of this chapter will detail the ADM in Poisson, normal and binomial models. For the multinomial model, please refer to Chapter 3 of this dissertation.

2.2 Poisson

Christiansen and Morris[13] propose the procedure PRIMM which takes advantage of ADM approximation when estimating the random individual Poisson parameters in a hierarchical Poisson-Gamma regression model. In their data example, this procedure provides shrinkage for individual data points and takes the severity of case mix in an individual hospital into account when computing the mortality rate in hospitals[14]. There are two mathematically equivalent methods to describe the hierarchical Poisson-gamma regression model, which are called the descriptive model and the inferential model, respectively. The two models are mathematically equivalent in the sense that they generate the same joint distribution for the data and the individual Poisson parameters conditional on the hyper-parameters.

There are three levels in the descriptive model. Level 1 concerns the distribution of each observation $\{(z_i, e_i)\}$, $i = 1, \ldots, N$, given the individual parameter $\{\lambda_i\}$, $i = 1, \ldots, N$. The notation z_i is the individual count and the notation e_i is the individual exposure. Level 2 gives the distribution of the individual parameter $\{\lambda_i\}$, $i = 1, \ldots, N$, conditional on the hyper-parameters $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{r-1}) \in \mathbb{R}^r$ and ζ. Level 3 concerns the hyper-prior distribution

of the hyper-parameters. The terms in the bracket notation $[,]$ are the mean and the variance for each distribution.

Level 1: Individual Model.

$$z_i|\lambda_i \sim Pois(e_i\lambda_i) = Pois[e_i\lambda_i, e_i\lambda_i], \tag{2.5}$$

independently, $i = 1, \ldots, N$.

Level 2: Structural Model.

$$\lambda_i|\boldsymbol{\beta}, \zeta \sim Gam(\zeta, \zeta/\mu_i) = Gam[\mu_i, \mu_i^2/\zeta],\ \log(\mu_i) = \boldsymbol{x_i'}\boldsymbol{\beta},\ \zeta > 0, \tag{2.6}$$

where $\boldsymbol{x_i'} \in \boldsymbol{R}^r$ and $\boldsymbol{\beta} \in \boldsymbol{R}^r$.

Level 3: Distributions of the Structural Parameters. This hyper-prior approximates a uniform distribution on the shrinkage.

$$h(\boldsymbol{\beta}, \zeta) = \frac{z_0}{(\zeta + z_0)^2}. \tag{2.7}$$

Equivalently,

$$B_0 = \frac{\zeta}{\zeta + z_0} \sim Uniform(0, 1). \tag{2.8}$$

There are also three levels in the inferential model. Level 1 in the inferential model is the Gamma mixture (2.6) of the Poisson distributions (2.5) in the descriptive model. And the distribution for the observation z_i conditional on the

hyper-parameters $\boldsymbol{\beta}$, ζ happens to be Negative Binomial. B_i is the shrinkage for each observation, which approaches 1 when the exposure e_i approaches 0 and approaches 0 when e_i approaches ∞. Level 2 here is the posterior distribution for the individual parameters λ_i conditional on the hyper-parameters. Since the gamma distribution is the conjugate prior of the Poisson distribution, the posterior for $\{\lambda_i\}$ is still a gamma distribution, but with updated parameters. As seen from the expression of λ_i^*, when e_i is large, the λ_i^* shrinks toward the observation and vice versa. Level 3 is the same as in the descriptive model.

Level 1: Marginal model for the observations.

$$z_i|\boldsymbol{\beta}, \zeta \sim NB(\zeta, 1 - B_i) = NB[e_i\mu_i, e_i\mu_i/B_i], \tag{2.9}$$

where

$$B_i = \frac{\zeta}{\zeta + e_i\mu_i}. \tag{2.10}$$

Level 2: Conditional model for the individual parameters.

$$\lambda_i|data, \boldsymbol{\beta}, \zeta \sim Gam(z_i + \zeta, e_i + \zeta/\mu_i) = Gam[\lambda_i^*, (\sigma_i^*)^2], \tag{2.11}$$

where

$$\lambda_i^* = E(\lambda_i|data, \boldsymbol{\beta}, \zeta) = (1 - B_i)y_i + B_i\mu_i, \tag{2.12}$$

and

$$(\sigma_i^*)^2 = Var(\lambda_i|data, \boldsymbol{\beta}, \zeta) = \lambda_i^*(1 - B_i)/e_i. \tag{2.13}$$

The distribution of the individual parameter $\lambda_i|data$ is of ultimate interest. In Level 2 of the inferential model, the conditional distribution of $\lambda_i|data, \boldsymbol{\beta}, \zeta$ is given. Combined with the joint distribution of the hyper-parameters $\boldsymbol{\beta}$, ζ in Level 3, it is possible to get the distribution of $\lambda_i|data$ by integration with respect to the hyper-parameters $\boldsymbol{\beta}, \zeta$. But this is computationally cumbersome. To save the trouble of taking integrals while still making accurate estimation, one important step in the procedure is to provide more accurate estimates for the hyper-parameters $\boldsymbol{\beta}$ and ζ. There are two main reasons for this: (1) $\lambda_i|data$ can be approximated properly by gamma distribution once the estimates for the hyper-parameters $\boldsymbol{\beta}$, ζ are accurate enough because $\lambda_i|data, \boldsymbol{\beta}, \zeta$ is gamma distributed; (2) the first two moments $E(\lambda_i|data) = E(\lambda_i^*|data)$ and $Var(\lambda_i|data) = Var(\lambda_i^*|data) + E((\sigma_i^*)^2|data)$ of the gamma distribution for $\lambda_i|data$ depends on the hyper-parameters $\boldsymbol{\beta}$, ζ.

For the purpose of getting accurate estimates for the hyper-parameters and avoiding an improper posterior, there are some adjustments made in the paper. The first adjustment is to introduce a proper distribution for ζ in Level 3. This adjustment is to prevent improper posterior for the hyper-parameter ζ. Without this adjustment, the likelihood for $\boldsymbol{\beta}, \zeta$ directly from Level 1 in the

inferential model is

$$L(\boldsymbol{\beta}, \zeta) = \prod_{i=1}^{k} \frac{\Gamma(\zeta + z_i)}{\Gamma(\zeta) z_i!} (1 - B_i)^{z_i} B_i^{\zeta}. \tag{2.14}$$

The regular maximum likelihood estimate (MLE) for ζ using $L(\boldsymbol{\beta}, \zeta)$ can occur at infinity, which will cause problems in inference. Through this adjustment, the posterior density for $\boldsymbol{\beta}$, ζ is

$$p(\boldsymbol{\beta}, \zeta) = c_0 L(\boldsymbol{\beta}, \zeta) z_0 / (\zeta + z_0)^2. \tag{2.15}$$

It can be shown that $p(\boldsymbol{\beta}, \zeta)$ is proper for both $\boldsymbol{\beta}$ and ζ provided that $N - n_0 \geq r$ and the $(N - n_0) \times r$ sub-matrix of non-zero count groups of $\boldsymbol{X}$ is of full rank, where n_0 is the number of observations with zero counts. Once the mild condition of the data is satisfied, the mode will occur at finite values for both $\boldsymbol{\beta}$ and ζ.

A second adjustment is to apply a restricted maximum likelihood (REML) type correction[19] to the posterior $p(\boldsymbol{\beta}, \zeta)$ with transformation of variable by setting $\tau = \log(\zeta)$,

$$p_2(\tau) = c_2 |\hat{\boldsymbol{H}}_{\boldsymbol{\tau}}|^{-1/2} L(\hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}}, e^{\tau}) e^{\tau} / (e^{\tau} + z_0)^2. \tag{2.16}$$

In $p_2(\tau)$, $\hat{\boldsymbol{H}}_{\boldsymbol{\tau}}$ is the second derivative of $\mathscr{L}(\boldsymbol{\beta}, \zeta) = log\Big(L(\boldsymbol{\beta}, \zeta)\Big)$ with respect

to $\boldsymbol{\beta}$ evaluated at regular MLE for $\boldsymbol{\beta}$ with fixed τ,

$$\hat{\boldsymbol{H}}_{\boldsymbol{\tau}} = -\frac{\partial^2 \mathscr{L}(\boldsymbol{\beta}, \zeta)}{\partial(\boldsymbol{\beta}\boldsymbol{\beta}')}\bigg|_{\hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}}} = \boldsymbol{X'} \boldsymbol{D}_{\boldsymbol{\tau}} \boldsymbol{X}. \tag{2.17}$$

In (2.17), $\hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}}$ is the unique solution to $\frac{\partial \mathscr{L}(\boldsymbol{\beta},\zeta)}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{k}(z_i - e_i\mu_i)B_i\boldsymbol{x_i}$ with τ fixed and $\hat{\boldsymbol{D}}_{\boldsymbol{\tau}}$ is the $N \times N$ diagonal matrix with diagonal element $e_i B_i \lambda_i^* > 0$, with B_i in (2.10) and λ_i^* in (2.12). This adjustment is to eliminate the bias in the estimate of the hyper-parameter ζ caused by the estimation of the hyper-parameter $\boldsymbol{\beta}$.

A third adjustment is to approximate $|\hat{\boldsymbol{H}}_{\boldsymbol{\tau}}|^{-1/r}$ by a constant multiplied by the geometric mean of the N values $\zeta/(\zeta + e_i m_0)$ with $m_0 = E(\sum z_i / \sum e_i)$. This adjustment simplifies the computation significantly and speeds up the procedure. Then the natural log of the adjusted posterior $p_2(\tau)$ is equal to

$$\begin{aligned} l_R(\boldsymbol{\beta}, \tau) &= \log\left\{ L(\boldsymbol{\beta}, \zeta | data) \frac{e^\tau}{(e^\tau + z_0)^2} \prod_{i=1}^{k} (e^\tau/(e^\tau + e_i m_0))^{-r/2k} \right\} \\ &= \mathscr{L}(\boldsymbol{\beta}, \zeta) + (1 - r/2)\tau - 2log\Big(exp(\tau) + z_0\Big) \\ &\quad + \frac{r}{2k}\sum_{i=1}^{k} log\Big(exp(\tau) + e_i m_0\Big). \end{aligned} \tag{2.18}$$

After the adjustments, take the first and second derivatives of $l_R(\boldsymbol{\beta}, \tau)$ with respect to $\boldsymbol{\beta}, \tau$. It is obvious that the first and second derivatives with respect to $\boldsymbol{\beta}$ and the second cross-derivative are the same as in the likelihood case. It follows that $\hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}}$ does not change after the adjustments. The first and second

derivatives of $l_R(\boldsymbol{\beta}, \tau)$ with respect to $\tau = \log(\zeta)$ are

$$\frac{\partial l_R(\boldsymbol{\beta}, \tau)}{\partial \tau} = \frac{\partial \mathscr{L}(\boldsymbol{\beta}, \zeta)}{\partial \tau} - \frac{e^\tau - z_0}{e^\tau + z_0} - \frac{r}{2k} \sum_{i=1}^{k} \frac{e_i m_0}{e^\tau + e_i m_0}, \tag{2.19}$$

and

$$-\frac{\partial^2 l_R(\boldsymbol{\beta}, \zeta)}{\partial \tau^2} = -\frac{\partial^2 \mathscr{L}(\boldsymbol{\beta}, \zeta)}{\partial \tau^2} + \frac{2 z_0 e^\tau}{(e^\tau + z_0)^2} - \frac{r}{2k} \sum_{i=1}^{k} \frac{e_i m_0 e^\tau}{(e^\tau + e_i m_0)^2}. \tag{2.20}$$

By inserting $\hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}}$, Newton's method can be used to find the zero of (2.19) using both (2.19) and (2.20). The zero of (2.19) denoted by $\hat{\tau}$ is the estimate of τ and evaluate $\hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}}$ at $\hat{\tau}$ to get the estimate for $\boldsymbol{\beta}$. The covariance matrix for $(\boldsymbol{\beta}, \tau)$ can be estimated using the inverse of the Hessian matrix of $l_R(\boldsymbol{\beta}, \tau)$ evaluated at the estimates $(\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\tau}}}, \hat{\tau})$. After some computation, the approximate distribution for $(\boldsymbol{\beta}, \tau)$ is multivariate normal:

$$\begin{pmatrix} \boldsymbol{\beta} \\ \tau \end{pmatrix} | data \sim N_{r+1} \left[\begin{pmatrix} \hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\tau}}} \\ \hat{\tau} \end{pmatrix}, \boldsymbol{\Sigma} = \hat{\sigma}_\tau^2 \begin{pmatrix} \hat{\sigma}_\tau^{-2} (\boldsymbol{X'} \hat{\boldsymbol{D}}_{\hat{\boldsymbol{\tau}}} \boldsymbol{X})^{-1} - \hat{\boldsymbol{v}} \hat{\boldsymbol{v}}' & \hat{\boldsymbol{v}} \\ \hat{\boldsymbol{v}} & 1 \end{pmatrix} \right]. \tag{2.21}$$

Here, $\hat{\boldsymbol{v}} = \partial \hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}} / \partial \tau$ and $\hat{\sigma}_\tau^2 = Var(\tau | data) \approx (-\partial^2 l_R(\boldsymbol{\beta}, \tau) / \partial \tau^2 - \hat{\boldsymbol{v}}' (\boldsymbol{X'} \hat{\boldsymbol{D}}_{\hat{\boldsymbol{\tau}}} \boldsymbol{X}) \hat{\boldsymbol{v}})^{-1}$.

The next step is to use the distribution (2.21) and ADM approximation to the distribution of the shrinkages $B_i | data$ to get a distribution for the individual parameter λ_i conditional on the data. Assume the distribution (2.21) is accurate enough so that treating $\lambda_i | data$ as Gamma distributed is appropri-

ate. Then the problem left is to estimate the first and second moments of the Gamma distribution. The posterior mean for λ_i is

$$\begin{aligned}\hat{\lambda}_i &= E(\lambda_i|data) = E_{\boldsymbol{\beta},\tau}[E(\lambda_i|data,\boldsymbol{\beta},\tau)|data] \\ &= E_{\boldsymbol{\beta},\tau}[(1-B_i)y_i + B_i\mu_i|data] \\ &= \Big(1 - E_{\boldsymbol{\beta},\tau}(B_i|data)\Big)y_i + E_{\boldsymbol{\beta},\tau}(B_i\mu_i|data).\end{aligned} \tag{2.22}$$

The posterior variance for λ_i is

$$\begin{aligned}\hat{\sigma}^2_{\lambda_i} &= Var(\lambda_i|data) \\ &= Var_{\boldsymbol{\beta},\tau}[E(\lambda_i|data,\boldsymbol{\beta},\tau)|data] + E_{\boldsymbol{\beta},\tau}[Var(\lambda_i|data,\boldsymbol{\beta},\tau)|data] \\ &= Var_{\boldsymbol{\beta},\tau}[(1-B_i)y_i + B_i\mu_i|data] \\ &\quad + E_{\boldsymbol{\beta},\tau}([(1-B_i)y_i + B_i\mu_i][1-B_i]/e_i|data) \\ &= E_{\boldsymbol{\beta},\tau}([(1-B_i)y_i + B_i\mu_i]^2|data) \\ &\quad + E_{\boldsymbol{\beta},\tau}[(1-B_i)^2y_i + B_i(1-B_i)\mu_i|data]/e_i - \hat{\lambda}_i^2.\end{aligned} \tag{2.23}$$

The subscript β and τ in (2.22) and (2.23) means the expectation and variance are with respect to the posterior distribution of β and τ. Both (2.22) and (2.23) are functions of $E_{\boldsymbol{\beta},\tau}(\mu_i^s B_i^t|data), s,t = 0,1,2$. Hereafter, to simplify notation, E stands for $E_{\boldsymbol{\beta},\tau}$ if not specified. Introduce the notation $E_sB_i^t$ and it can be proved that

$$E_sB_i^t = \frac{E\mu_i^sB_i^t}{E\mu_i^s} = \frac{EB_i^t(B_i/(1-B_i))^{sb_i}}{E(B_i/(1-B_i))^{sb_i}}, \tag{2.24}$$

with $b_i = Cov(\boldsymbol{x_i'\beta}, \tau - \boldsymbol{x_i'\beta})/Var(\tau - \boldsymbol{x_i'\beta})$. This equation successfully trans-

forms the two-dimensional expectation into a one-dimensional expectation. $E\mu_i^s$ can be easily obtained from the moment generating function (MGF) of normal distribution. Then, if the right side of (2.24) can be approximated, $E\mu_i^s B_i^t$ can be computed quickly without doing integrals. ADM approximation[37] can be used to approximate the right side of (2.24). Since B_i has support $(0,1)$, it is proper to choose beta distribution to do the approximation. $B_i = \frac{\zeta}{\zeta + e_i\mu_i} = \frac{e^{\tau}}{e^{\tau} + e_i e^{\boldsymbol{x_i'\beta}}} = \frac{e^{\tau - \boldsymbol{x_i'\beta}}}{e^{\tau - \boldsymbol{x_i'\beta}} + e_i} = \frac{e^{\tau - \boldsymbol{x_i'\beta}}}{e^{\tau - \boldsymbol{x_i'\beta}} + e^{\ln(e_i)}} = \frac{e^{\tau - \boldsymbol{x_i'\beta} - \ln(e_i)}}{1 + e^{\tau - \boldsymbol{x_i'\beta} - \ln(e_i)}}$ and denote $u = \tau - \boldsymbol{x_i'\beta} - ln(e_i) \sim N(\mu, \sigma^2)$ from (2.21). Drop the subscript i for B_i since the procedure will be the same for all i. Assume $B = \frac{e^u}{1+e^u}$, then B is logit-normal distribution with density $p(B)$. Now use the beta distribution to approximate the logit-normal distribution by the following procedures.

$$B \sim Beta(a_1, a_2) = Beta\left[\hat{B} = \frac{a_1}{a_1 + a_2}, \frac{\hat{B}(1-\hat{B})}{a_1 + a_2 + 1}\right]. \tag{2.25}$$

Give the expression for $l(B) = log(p(B)B(1-B))$ with $B(1-B)$ being the adjustment factor for beta approximation. Compute $\hat{B}$ which maximizes $l(B)$ and the second derivative of $l(\hat{B})$.

$$l(B) = log(logit - normal(B)B(1-B)) \propto (logit(B) - \mu)^2 \tag{2.26}$$

$$\frac{\partial l(B)}{\partial B} \propto \frac{logit(B) - \mu}{B(1-B)} = 0 \Rightarrow \hat{B} = \frac{e^{\mu}}{1 + e^{\mu}} \tag{2.27}$$

$$-\ddot{l}(\hat{B}) = -\left.\frac{\partial^2 l(B)}{\partial B^2}\right|_{\hat{B}} = \frac{1}{\hat{B}^2(1-\hat{B})^2\sigma^2} \tag{2.28}$$

Then solve for the system of equations to get the expressions for the approximation parameters $\hat{a}_1$ and $\hat{a}_2$.

$$\begin{aligned} \frac{\hat{a}_1}{\hat{a}_1+\hat{a}_2} &= \hat{B} \\ \hat{a}_1+\hat{a}_2 &= \frac{1}{\hat{B}(1-\hat{B})\sigma^2}. \end{aligned} \tag{2.29}$$

The solution $\hat{a}_1$ and $\hat{a}_2$ are

$$\hat{a}_1 = \frac{1}{\sigma^2(1-\hat{B})} \tag{2.30}$$

and

$$\hat{a}_2 = \frac{1}{\sigma^2\hat{B}} \tag{2.31}$$

Now B_i can be approximated by $Beta(\hat{a}_1, \hat{a}_2)$. And the right-hand side of equation (2.24) can be computed fast by

$$\begin{aligned} &\frac{EB^t(B/(1-B))^w}{E(B/(1-B))^w} = \frac{\int B^t(B/(1-B))^w \frac{\Gamma(a_1+a_2)}{\Gamma(a_1)\Gamma(a_2)} B^{a_1-1}(1-B)^{a_2-1}dB}{\int (B/(1-B))^w \frac{\Gamma(a_1+a_2)}{\Gamma(a_1)\Gamma(a_2)} B^{a_1-1}(1-B)^{a_2-1}dB} \\ =&\frac{\int B^{t+w+a_1-1}(1-B)^{a_2-w-1}dB}{\int B^{w+a_1-1}(1-B)^{a_2-w-1}dB} \\ =&\frac{\Gamma(t+w+a_1)\Gamma(a_2-w)}{\Gamma(t+a_1+a_2)}\frac{\Gamma(a_1+a_2)}{\Gamma(w+a_1)\Gamma(a_2-w)}\frac{\int Beta(t+w+a_1, a_2-w)dB}{\int Beta(w+a_1, a_2-w)dB} \\ =&\frac{\Gamma(t+w+a_1)}{\Gamma(w+a_1)}\frac{\Gamma(a_1+a_2)}{\Gamma(t+a_1+a_2)} = \frac{(t+w+a_1-1)...(w+a_1)}{(t+a_1+a_2-1)...(a_1+a_2)} \end{aligned} \tag{2.32}$$

This approximation simplifies and speeds up the computation of $E\mu_i^s B_i^t$ and correspondingly, the computation of the posterior mean in equation (2.22) and

the posterior variance in equation (2.23). And the posterior distribution for λ_i can be determined as

$$\lambda_i|data \dot{\sim} Gam[\hat{\lambda}_i, \hat{\sigma}^2_{\lambda_i}] = \frac{\hat{\sigma}^2_{\lambda_i}}{\hat{\lambda}_i} Gam\left(\frac{\hat{\lambda}^2_i}{\hat{\sigma}^2_{\lambda_i}}, 1\right). \tag{2.33}$$

2.3 Normal

Researches in ADM approximation to normal-normal model were mainly conducted by Tang and Morris[38]. Kelly and Morris[24] add the ADM approximation to skewness in the normal model to this collection of literature.

There are two levels in the normal-normal hierarchical model. As in the Poisson case, there are also two mathematically equivalent models, the descriptive model and the inferential model, in this normal case. In the descriptive model, the first level specifies the individual normal distribution of the observed data y_i, given the individual parameters θ_i, $i = 1, \ldots, N$. Level 2 specifies the normal distributions of θ_i, $i = 1, \ldots, N$, given the hyper-parameters $\boldsymbol{\beta}$ and A. The hyper-parameters $\boldsymbol{\beta} \in \mathbb{R}^m$ are the regression coefficients.

Level 1: The individual observations y_i conditional on the individual parameters θ_i are independently normal with unknown mean θ_i and known variance V_i, $i = 1, \ldots, N$:

$$y_i|\theta_i \sim N(\theta_i, V_i). \tag{2.34}$$

Level 2: The individual parameters θ_i, $i = 1, \ldots, N$, given the unknown hyper-parameters $\boldsymbol{\beta}$ and A, are also independently normally distributed:

$$\theta_i | \boldsymbol{\beta}, A \sim N(\boldsymbol{x}_i' \boldsymbol{\beta}, A), \tag{2.35}$$

in which $\boldsymbol{x}_i \in \mathbb{R}^m$ is known, $\boldsymbol{\beta} \in \mathbb{R}^m$ is unknown and A is an unknown scalar.

The inferential model also has two levels. Level 1 is derived by integrating the individual parameter θ_i out. Level 2 is the posterior distribution of the individual parameter θ_i conditional on the hyper-parameters $\boldsymbol{\beta}$ and A.

Level 1: This level is known as the marginal distribution of the data y_i conditional on the hyper-parameters $\boldsymbol{\beta}$ and A. The marginal distribution in this level is still a normal distribution for each observation y_i, $i = 1, \ldots, N$,

$$y_i | \boldsymbol{\beta}, A \sim N(\boldsymbol{x}_i' \boldsymbol{\beta}, V_i + A), \tag{2.36}$$

independently.

Level 2: Posterior distributions of the individual parameters θ_i given the hyper-parameters $\boldsymbol{\beta}$ and A are independent normal distributions.

$$\theta_i | y_i, \boldsymbol{\beta}, A \sim N\Big((1 - B_i) y_i + B_i \boldsymbol{x}_i' \boldsymbol{\beta}, V_i (1 - B_i)\Big), \tag{2.37}$$

where $B_i = V_i / (V_i + A)$ is the shrinkage factor and the conditional posterior

mean and conditional posterior variance are

$$\theta_i^* = E(\theta_i|y_i, \boldsymbol{\beta}, A) = (1 - B_i)y_i + B_i\boldsymbol{x}_i'\boldsymbol{\beta}, \tag{2.38}$$

and

$$Var(\theta_i|y_i, \boldsymbol{\beta}, A) = V_i(1 - B_i), \tag{2.39}$$

respectively. In the normal-normal model, the hyper-priors for the hyper-parameters are flat for both $\boldsymbol{\beta}$ and A. That is,

$$f(\boldsymbol{\beta}, A) \propto 1. \tag{2.40}$$

The likelihood of the hyper-parameters $\boldsymbol{\beta}$ and A is

$$\begin{aligned} L(\boldsymbol{\beta}, A) &= \prod_{i=1}^{N} N(y_i|\boldsymbol{x}_i'\boldsymbol{\beta}, V_i + A) \\ &\propto \exp\Big(-\sum_{i=1}^{N} \frac{(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2(V_i + A)} \Big) \prod_{i=1}^{N} (V_i + A)^{-\frac{1}{2}} \end{aligned} \tag{2.41}$$

This likelihood can also be written in matrix-vector notation

$$L(\boldsymbol{\beta}, A) \propto |\boldsymbol{D}_{V+A}|^{-1/2} \exp(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{D}_{V+A}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})), \tag{2.42}$$

where $\boldsymbol{D}_{V+A}$ is the $N \times N$ diagonal matrix with diagonal terms $V_i + A$. Since the hyper-prior of the hyper-parameters is flat, $f(\boldsymbol{\beta}, A) \propto 1$, the posterior of

$\boldsymbol{\beta}$ and A is equal to the likelihood of $\boldsymbol{\beta}$ and A,

$$p(\boldsymbol{\beta}, A|\boldsymbol{y}) \propto L(\boldsymbol{\beta}, A) f(\boldsymbol{\beta}, A) = L(\boldsymbol{\beta}, A). \tag{2.43}$$

It has been proved by Kelly[24] that the sufficient data-dependent posterior propriety condition is that $N \geq m + 3$ where N is the number of observations in the sample and m is the rank of the $N \times m$ covariate matrix $\boldsymbol{X}$.

There are two cases in estimating the random effects θ_i in the normal model, which are (1) equal variances with all the $V_i = V$ for $i = 1, \ldots, N$ and (2) unequal variances. It is possible to get exact moments for the shrinkage factors B_i and the exact means and variances for the individual parameters θ_i in the first case. Then the ADM approximation is only applied to the second case.

Before moving on, they derive two useful distributions which will be used later. The first is the conditional posterior distribution for $\boldsymbol{\beta}$:

$$\begin{aligned} p(\boldsymbol{\beta}|\boldsymbol{y}, A) &\propto \exp\Big(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})' \boldsymbol{D}_{V+A}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \Big) \\ &= \exp\Bigg(-\frac{1}{2}\Big((\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_A) + \boldsymbol{X}(\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}) \Big)' \\ &\qquad \boldsymbol{D}_{V+A}^{-1} \Big((\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_A) + \boldsymbol{X}(\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}) \Big) \Bigg) \\ &\propto \exp\Big(-\frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_A)' \boldsymbol{\Sigma}_A^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_A) \Big), \end{aligned} \tag{2.44}$$

where $\boldsymbol{\Sigma}_A = \Big(\boldsymbol{X}' \boldsymbol{D}_{V+A}^{-1} \boldsymbol{X} \Big)^{-1}$ and $\hat{\boldsymbol{\beta}}_A$ is the weighted least squares estimator

for $\boldsymbol{\beta}$: $\hat{\boldsymbol{\beta}}_A = \boldsymbol{\Sigma_A X' D}_{V+A}^{-1}\boldsymbol{y}$. In matrix-vector notation

$$\boldsymbol{\beta}|\boldsymbol{y}, A \sim N_m(\hat{\boldsymbol{\beta}}_A, \boldsymbol{\Sigma}_A). \tag{2.45}$$

The other useful distribution is the posterior distribution for the hyper-parameter A:

$$\begin{aligned} p(A|\boldsymbol{y}) &= \int_{\boldsymbol{\beta}} p(\boldsymbol{\beta}, A|\boldsymbol{y}) d\boldsymbol{\beta} \\ &\propto \int_{\boldsymbol{\beta}} |\boldsymbol{D}_{V+A}|^{-1/2} exp\Big(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X\beta})'\boldsymbol{D}_{V+A}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})\Big) d\boldsymbol{\beta} \\ &= |\boldsymbol{D}_{V+A}|^{-1/2}|\boldsymbol{\Sigma}_A|^{1/2} exp\Big(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_A)'\boldsymbol{D}_{V+A}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_A)\Big). \end{aligned} \tag{2.46}$$

Under the condition of equal variances $V_i = V$, equation (2.45) reduces to

$$\boldsymbol{\beta}|\boldsymbol{y}, A \sim N(\hat{\boldsymbol{\beta}}, (V+A)(\boldsymbol{X'X})^{-1}), \tag{2.47}$$

where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$ is the traditional least squares estimator. Then equation (2.46) reduces to

$$p(A|\boldsymbol{y}) \propto (V+A)^{-(N-m)/2} e^{-S/(2(V+A))}, \tag{2.48}$$

where $S = \sum_{i=1}^{N}(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}})^2$. The conditional posterior distribution of the random effect θ_i in equation (2.37) reduces to

$$\theta_i|\boldsymbol{y}, \boldsymbol{\beta}, A \sim\sim N\Big((1-B)y_i + B\boldsymbol{x}_i'\boldsymbol{\beta}, V(1-B)\Big), \tag{2.49}$$

where $B = V/(V+A)$. The objective is to estimate the posterior mean $E(\theta_i|\boldsymbol{y})$ and posterior variance $Var(\theta_i|\boldsymbol{y})$ of the individual parameters θ_i. Apply the Law of Total Expectation and the Law of Total Variance first over $\boldsymbol{\beta}$,

$$\begin{aligned} E(\theta_i|A,\boldsymbol{y}) &= E(\theta_i^*|A,\boldsymbol{y}) = E\Big((1-B)y_i + B\boldsymbol{x_i'}\boldsymbol{\beta}|A,\boldsymbol{y}\Big) \\ &= (1-B)y_i + B\boldsymbol{x_i'}\hat{\boldsymbol{\beta}} \end{aligned} \tag{2.50}$$

where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$ and

$$\begin{aligned} Var(\theta_i|A,\boldsymbol{y}) &= Var(\theta_i^*|A,\boldsymbol{y}) + E\Big(V(1-B)|A,\boldsymbol{y}\Big) \\ &= Var\Big(B\boldsymbol{x_i'}\boldsymbol{\beta}|A,\boldsymbol{y}\Big) + V(1-B) \\ &= B^2(V+A)\boldsymbol{x_i'}(\boldsymbol{X'X})^{-1}\boldsymbol{x}_i + V(1-B) \\ &= BV\boldsymbol{x_i'}(\boldsymbol{X'X})^{-1}\boldsymbol{x}_i + V(1-B). \end{aligned} \tag{2.51}$$

Then apply the Law of Total Expectation and the Law of Total Variance over A,

$$\begin{aligned} E(\theta_i|\boldsymbol{y}) &= E((1-B)y_i + B\boldsymbol{x_i'}\hat{\boldsymbol{\beta}}|\boldsymbol{y}) \\ &= \Big(1 - E(B|\boldsymbol{y})\Big)y_i + E(B|\boldsymbol{y})\boldsymbol{x_i'}\hat{\boldsymbol{\beta}}, \end{aligned} \tag{2.52}$$

and

$$\begin{aligned} Var(\theta_i|\boldsymbol{y}) &= Var\Big((1-B)y_i + B\boldsymbol{x_i'}\hat{\boldsymbol{\beta}}|\boldsymbol{y}\Big) \\ &\quad + E\Big(BV\boldsymbol{x_i'}(\boldsymbol{X'X})^{-1}\boldsymbol{x}_i + V(1-B)|\boldsymbol{y}\Big) \\ &= (y_i - \boldsymbol{x_i'}\hat{\boldsymbol{\beta}})^2 Var(B|\boldsymbol{y}) \\ &\quad + V\boldsymbol{x_i'}(\boldsymbol{X'X})^{-1}\boldsymbol{x}_i E(B|\boldsymbol{y}) + V\Big(1 - E(B|\boldsymbol{y})\Big). \end{aligned} \tag{2.53}$$

Then the remaining problem is to estimate $E(B|\boldsymbol{y})$ and $Var(B|\boldsymbol{y})$. Since the posterior distribution of A has been given in equation (2.48) and $B = V/(V+A)$ is a function of A, they perform a variable transformation. Then the posterior distribution for B is

$$\begin{aligned} p(B|\boldsymbol{y}) &\propto \left(\frac{V}{B}\right)^{-(N-m)/2} \exp\left(-\frac{SB}{2V}\right)\left(\frac{V}{B^2}\right) I\{0<B<1\} \\ &\propto B^{(N-m-2)/2-1} exp\left(-\frac{SB}{2V}\right) I\{0<B<1\}. \end{aligned} \tag{2.54}$$

From equation (2.54), the exact posterior distribution of B is actually a gamma distribution restricted to $0 < B < 1$:

$$B|\boldsymbol{y} \sim Gamma\Big(\frac{N-m-2}{2}, \frac{S}{2V}\Big), \tag{2.55}$$

where $0 < B < 1$. Denote $a = (N-m-2)/2$ and $b = S/2V$. Then, the exact posterior moment of B for any power c is

$$\begin{aligned} E(B^c|\boldsymbol{y}) &= \frac{\int_0^1 B^c b^a (\Gamma(a))^{-1} B^{a-1} e^{-bB} dB}{P(G_{a,b}<1)} \\ &= b^{-c}\frac{\Gamma(a+c)}{\Gamma(a)}\frac{P(G_{a+c,b}<1)}{P(G_{a,b}<1)}. \end{aligned} \tag{2.56}$$

From equation (2.56), the exact posterior mean $E(B|\boldsymbol{y})$ and posterior variance $Var(B|\boldsymbol{y})$ are

$$\begin{aligned} E(B|\boldsymbol{y}) &= ab^{-1}\frac{P(G_{a+1,b}<1)}{P(G_{a,b}<1)} \\ &= \frac{(N-m-2)V}{S}\frac{P(\chi^2_{N-m}<S/V)}{P(\chi^2_{N-m-2}<S/V)}, \end{aligned} \tag{2.57}$$

and

$$
\begin{aligned}
Var(B|\boldsymbol{y}) &= E(B^2|\boldsymbol{y}) - \Big(E(B|\boldsymbol{y})\Big)^2 \\
&= (a+1)ab^{-2}\frac{P(G_{a+2,b}<1)}{P(G_{a,b}<1)} - \Big(E(B|\boldsymbol{y})\Big)^2 \\
&= \frac{N-m}{2}\frac{N-m-2}{2}(\frac{S}{2V})^{-2}\frac{P(\chi^2_{N-m+2}<S/V)}{P(\chi^2_{N-m-2}<S/V)} - \Big(E(B|\boldsymbol{y})\Big)^2 \\
&= (N-m)(N-m-2)(\frac{V^2}{S^2})\frac{P(\chi^2_{N-m+2}<S/V)}{P(\chi^2_{N-m-2}<S/V)} - \Big(E(B|\boldsymbol{y})\Big)^2,
\end{aligned}
\tag{2.58}
$$

respectively. Inserting the results from equation (2.57) and equation (2.58) into equation (2.52) and equation (2.53), the posterior mean $E(\theta_i|\boldsymbol{y})$ and the posterior variance $Var(\theta_i|\boldsymbol{y})$ for the individual parameters are obtained under the equal variance condition.

For the unequal variance condition, there are no exact values for the posterior mean $E(\theta_i|\boldsymbol{y})$ and posterior variance $Var(\theta_i|\boldsymbol{y})$, but the ADM approximation can be applied to approximate $E(\theta_i|\boldsymbol{y})$ and $Var(\theta_i|\boldsymbol{y})$. Apply the Law of Total Expectation and the Law of Total Variance over $\boldsymbol{\beta}$

$$
\begin{aligned}
E(\theta_i|A,\boldsymbol{y}) = E(\theta_i^*|A,\boldsymbol{y}) &= E\Big((1-B_i)y_i + B_i\boldsymbol{x_i'}\boldsymbol{\beta}|A,\boldsymbol{y}\Big) \\
&= (1-B_i)y_i + B_i\boldsymbol{x_i'}\hat{\boldsymbol{\beta}}_A,
\end{aligned}
\tag{2.59}
$$

where $\hat{\boldsymbol{\beta}}_A = \Big(\boldsymbol{X'}\boldsymbol{D}_{V+A}^{-1}\boldsymbol{X}\Big)^{-1}\boldsymbol{X'}\boldsymbol{D}_{V+A}^{-1}\boldsymbol{y}$, and

$$\begin{aligned} Var(\theta_i|A,\boldsymbol{y}) &= Var(\theta_i^*|A,\boldsymbol{y}) + E\Big(V_i(1-B_i)|A,\boldsymbol{y}\Big) \\ &= Var\Big(B_i\boldsymbol{x}_i'\boldsymbol{\beta}|A,\boldsymbol{y}\Big) + V_i(1-B_i) \\ &= B_i^2\boldsymbol{x}_i'(\boldsymbol{X'}\boldsymbol{D}_{V+A}^{-1}\boldsymbol{X})^{-1}\boldsymbol{x}_i + V_i(1-B_i) \\ &= B_iV_i\Big[\boldsymbol{x}_i'(\boldsymbol{X'}\boldsymbol{D}_{V+A}^{-1}\boldsymbol{X})^{-1}\boldsymbol{x}_i)/(V_i+A)\Big] + V_i(1-B_i). \end{aligned} \tag{2.60}$$

Next, they evaluate $\hat{\boldsymbol{\beta}}_A$ and the quantity inside the bracket in the last line in equation (2.60) at the optimal $\hat{A}$ and then apply the Law of Total Expectation and the Law of Total Variance over A to obtain

$$\begin{aligned} \hat{\theta}_i = E(\theta_i|\boldsymbol{y}) &\approx E\Big((1-B_i)y_i + B_i\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{\hat{A}}|\boldsymbol{y}\Big) \\ &= \Big(1-E(B_i|\boldsymbol{y})\Big)y_i + E(B_i|\boldsymbol{y})\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{\hat{A}}, \end{aligned} \tag{2.61}$$

and

$$\begin{aligned} s_i = Var(\theta_i|\boldsymbol{y}) &\approx E\Big(B_iV_ip_{ii} + V_i(1-B_i)|\boldsymbol{y}\Big) + Var\Big((1-B_i)y_i + B_i\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{\hat{A}}|\boldsymbol{y}\Big) \\ &= E(B_i|\boldsymbol{y})V_ip_{ii} + V_i\Big(1-E(B_i|\boldsymbol{y})\Big) + (y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{\hat{A}})^2Var(B_i|\boldsymbol{y}) \\ &= \Big(1-(1-p_{ii})E(B_i|\boldsymbol{y})\Big)V_i + (y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{\hat{A}})^2Var(B_i|\boldsymbol{y}), \end{aligned} \tag{2.62}$$

where $p_{ii} = \boldsymbol{x}_i'(\boldsymbol{X'}\boldsymbol{D}_{V+\hat{A}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{x}_i)/(V_i+\hat{A})$.

There are two problems left to estimate $E(\theta_i|\boldsymbol{y})$ and $Var(\theta_i|\boldsymbol{y})$ for the unequal variance case. One is to find the optimal $\hat{A}$ and the other is to approximate the moments of $B_i|\boldsymbol{y}$. To find $\hat{A}$, first conduct a change of variable $\alpha = \log(A)$.

This transformation of variable is made because the distribution of α is more symmetric and has no boundary issues when applying MLE. The posterior distribution of α is

$$f(\alpha|\boldsymbol{y}) = f(A(\alpha)|\boldsymbol{y})e^{\alpha}, \tag{2.63}$$

and $\hat{\alpha} = argmax\Big(\alpha + log\Big(f(A(\alpha)|\boldsymbol{y})\Big)\Big)$. Set $\hat{A} = e^{\hat{\alpha}}$, and then approximate the distribution of $B_i|\boldsymbol{y}$ by a beta distribution because B_i is between 0 and 1:

$$B_i|\boldsymbol{y} \sim Beta(a_{i1}, a_{i2}). \tag{2.64}$$

. They use a similar procedure as used in the ADM in the Poisson model in Section 1.2. We have

$$\begin{aligned} B_i &= \frac{V_i}{V_i + A} = \frac{V_i}{V_i + e^{\alpha}} \\ &= \frac{V_i e^{-\alpha}}{1 + V_i e^{-\alpha}} \\ &= \frac{e^{-(\alpha - \log(V_i))}}{1 + e^{-(\alpha - \log(V_i))}} \sim logit - normal\Big(- \big(\hat{\alpha} - log(V_i)\big),\ \hat{\sigma}_{\alpha}^2\Big), \end{aligned} \tag{2.65}$$

where $\hat{\sigma}_{\alpha}^2$ can be approximated by the reciprocal of the Fisher information of $f(\alpha|\boldsymbol{y})$. Give the expression for $l(B_i) = \log\Big(p(B_i)B_i(1-B_i)\Big)$ with $B_i(1-B_i)$ being the adjustment factor for the beta approximation. Compute $\hat{B}_i$ which maximizes $l(B_i)$ and the second derivative of $l(B_i)$ evaluated at $\hat{B}_i$. Set $\mu_i =$

$-\Big(\hat{\alpha} - \log(V_i)\Big)$ to obtain:

$$l(B_i) = \log(logit - normal(B_i)B_i(1-B_i)) \propto (logit(B_i) - \mu_i)^2, \tag{2.66}$$

$$\frac{\partial l(B_i)}{\partial B_i} \propto \frac{logit(B_i) - \mu_i}{B_i(1-B_i)} = 0 \Rightarrow \hat{B}_i = \frac{e^{\mu_i}}{1+e^{\mu_i}} = \frac{V_i}{V_i + \hat{A}}, \tag{2.67}$$

$$-\ddot{l}(\hat{B}_i) = -\left.\frac{\partial^2 l(B_i)}{\partial B_i^2}\right|_{\hat{B}_i} = \frac{1}{\hat{B}_i^2(1-\hat{B}_i)^2\hat{\sigma}_\alpha^2}. \tag{2.68}$$

The system of equations for the approximation parameters $\hat{a}_{i1}$ and $\hat{a}_{i2}$ is

$$\begin{aligned} \frac{\hat{a}_{i1}}{\hat{a}_{i1}+\hat{a}_{i2}} &= \hat{B}_i, \\ \hat{a}_{i1}+\hat{a}_{i2} &= \frac{1}{\hat{B}_i(1-\hat{B}_i)\hat{\sigma}_\alpha^2}. \end{aligned} \tag{2.69}$$

The solution to the system of equations is

$$\hat{a}_{i1} = \frac{1}{\hat{\sigma}_\alpha^2(1-\hat{B}_i)}, \tag{2.70}$$

and

$$\hat{a}_{i2} = \frac{1}{\hat{\sigma}_\alpha^2\hat{B}_i}. \tag{2.71}$$

Then

$$\hat{B}_i = \hat{E}(B_i|\boldsymbol{y}) = \frac{V_i}{V_i + \hat{A}}, \tag{2.72}$$

and

$$\widehat{Var}(B_i|\boldsymbol{y}) = \frac{\hat{B}_i^2(1-\hat{B}_i)^2}{I_\alpha + \hat{B}_i(1-\hat{B}_i)}, \tag{2.73}$$

where I_α is the Fisher information of $f(\alpha|\boldsymbol{y})$. Inserting equation (2.72) and

equation (2.73) into the equation (2.61) and the equation (2.62), they obtain the mean and variance estimates of the random effect θ_i.

As mentioned at the beginning of this section, Kelly also applies ADM to estimate the skewness in a normal model. So far the first two central moments of $\theta_i|\boldsymbol{y}$, which are the posterior mean $\hat{\theta}_i$ and the posterior variance s_i, have been estimated. Since the skewness must be estimated, the third central moment of $\theta_i|\boldsymbol{y}$ must be estimated. The third central moment of $\theta_i|\boldsymbol{y}$ is computed as follows under the assumption that $\boldsymbol{\beta}_A$ has been estimated at $\hat{A}$.

$$\begin{aligned}
\mu_3(\theta_i|\boldsymbol{y}) &= E\Big(\mu_3(\theta_i|A,\boldsymbol{y})|\boldsymbol{y}\Big) + \mu_3\Big(E(\theta_i|A,\boldsymbol{y})|\boldsymbol{y}\Big) \\
&\quad + 3Cov\Big(E(\theta_i|A,\boldsymbol{y}), Var(\theta_i|A,\boldsymbol{y})|\boldsymbol{y}\Big) \\
&= 0 + \mu_3\Big((1-B_i)y_i + B_i\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{\hat{A}}|\boldsymbol{y}\Big) \\
&\quad + 3Cov\Big((1-B_i)y_i + B_i\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{\hat{A}}, V_i(1-B_i)|\boldsymbol{y}\Big) \\
&= -(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{\hat{A}})^3\mu_3(B_i|\boldsymbol{y}) + 3V_i(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{\hat{A}})Var(B_i|\boldsymbol{y}).
\end{aligned} \tag{2.74}$$

In equation (2.74), $Var(B_i|\boldsymbol{y})$ has been estimated in equation (2.73). Then estimate $\mu_3(B_i|\boldsymbol{y})$ by inserting the ADM estimates into the equation for the third central moment of the Beta distribution,

$$\hat{\mu}_3(B_i|\boldsymbol{y}) \approx \hat{a}_{i3} = \frac{2\hat{a}_{i1}\hat{a}_{i2}(\hat{a}_{i2}-\hat{a}_{i1})}{(\hat{a}_{i1}+\hat{a}_{i2})^3(\hat{a}_{i1}+\hat{a}_{i2}+2)\sqrt{\hat{a}_{i1}+\hat{a}_{i2}+1}}. \tag{2.75}$$

Inserting $\hat{\mu}_3(B_i|\boldsymbol{y})$ into equation (2.74) results in the estimate of $\mu_3(\theta_i|\boldsymbol{y})$, which is denoted by $\hat{\mu}_{3i}$ for notation simplicity.

For normal distribution with skewness, there are three parameters and is denoted by $skew - normal(\psi, \omega, \delta)$. Assume there is a random variable $Y \sim skew - normal(\psi, \omega, \delta)$. By matching the three central moments, it is possible to estimate the parameters ψ, ω, δ in the Skew-Normal distribution.

$$E(Y) = \psi_i + \omega_i \delta_i \sqrt{\frac{2}{\pi}} = \hat{\theta}_i, \tag{2.76}$$

$$Var(Y) = \omega_i^2 \left(1 - \frac{2\delta_i^2}{\pi}\right) = s_i, \tag{2.77}$$

$$skewness = \frac{4-\pi}{2} \frac{\delta_i^3}{(\pi/2 - \delta_i^2)^{3/2}} = \hat{\mu}_{3i}. \tag{2.78}$$

Solve the system of equations

$$\hat{\delta}_i = sign(\hat{\gamma}_i) \sqrt{\frac{\frac{\pi}{2}|\hat{\gamma}_i|^{2/3}}{|\hat{\gamma}_i|^{2/3} + ((4-\pi)/2)^{2/3}}}, \tag{2.79}$$

$$\hat{\omega}_i = \sqrt{\frac{s_i}{(1 - \frac{2\hat{\delta}_i^2}{\pi})}}, \tag{2.80}$$

$$\hat{\psi}_i = \hat{\theta}_i - \hat{\omega}_i \delta_i \sqrt{\frac{2}{\pi}}, \tag{2.81}$$

where $\hat{\gamma}_i = \hat{\mu}_{3i}/(s_i^{3/2})$.

2.4 Binomial

The ADM approximation in the binomial-beta-logit model is mainly documented in the work by Tak, Kelly and Morris[25]. There are three levels in

the binomial-beta-logit model. Level 1 specifies the binomial distribution for the individual observed data $\{(y_i, n_i)\}$, given the individual parameters p_i, $i = 1, \ldots, N$, where y_i is the number of success out of n_i trials. Level 2 specifies a beta distribution for the individual parameter p_i given the hyper-parameters $\boldsymbol{\beta} \in \mathbb{R}^m$ and r. A beta distribution is used because it is the conjugate prior for the binomial distribution. Level 3 assigns a hyper-prior to the hyper-parameters.

Level 1: The individual observations y_i given the individual parameters p_i are conditionally independent $binomial(n_i, p_i)$ distributions

$$y_i | p_i \sim Binomial(n_i,\ p_i). \tag{2.82}$$

Level 2: The individual parameters p_i given the hyper-parameters $\boldsymbol{\beta}$ and r have conditionally independent beta distributions

$$p_i | \boldsymbol{\beta}, r \sim Beta\Big(rp_i^E, r(1 - p_i^E)\Big), \tag{2.83}$$

where $p_i^E = E(p_i | \boldsymbol{\beta}, r) = e^{\boldsymbol{x}_i\boldsymbol{\beta}}/(1 + e^{\boldsymbol{x}_i\boldsymbol{\beta}})$.

Level 3: This level specifies the distribution for the hyper-parameters. The hyper-parameter $\boldsymbol{\beta}$ is assigned an improper flat distribution and the reciprocal of the hyper-parameter r is uniformly distributed on the interval $(0, \infty)$:

$$\boldsymbol{\beta} \sim Uniform\ on\ \boldsymbol{R}^m \tag{2.84}$$

and

$$1/r \sim Uniform(0, \infty). \tag{2.85}$$

The inferential model also contains three levels. The marginal distribution of the individual observations y_i given the hyper-parameters $\boldsymbol{\beta}$ and r are independent Beta-Binomial distributions with density:

$$f(y_i|\boldsymbol{\beta}, r) = \binom{n_i}{y_i} \frac{B(y_i + rp_i^E, n_i - y_i + r(1 - p_i^E))}{B(rp_i^E, r(1 - p_i^E))}, \tag{2.86}$$

where $B(a, b)(= \int_0^1 v^{a-1}(1 - v)^{b-1}dv)$ denotes the beta function for positive constants a and b.

Level 1: The individual observations y_i conditional on the hyper-parameters are independent beta-binomial distributions

$$y_i|\boldsymbol{\beta}, r \sim f(y_i|\boldsymbol{\beta}, r). \tag{2.87}$$

Level 2: The conditional posterior distributions of p_i given the hyper-parameters and the observed data $\boldsymbol{y}$ are conditionally independent beta distributions with updated parameters

$$p_i|\boldsymbol{\beta}, r, \boldsymbol{y} \sim Beta(n_i\bar{y}_i + rp_i^E, n_i(1 - \bar{y}_i) + r(1 - p_i^E)), \tag{2.88}$$

where $\bar{y}_i = y_i/n_i$. Thus, the conditional posterior mean and conditional pos-

terior variance of p_i are

$$p_i^* = E(p_i|\boldsymbol{\beta}, r, \boldsymbol{y}) = (1 - B_i)\bar{y}_i + B_i p_i^E \tag{2.89}$$

and

$$Var(p_i|\boldsymbol{\beta}, r, \boldsymbol{y}) = \frac{p_i^*(1 - p_i^*)}{r + n_i + 1}, \tag{2.90}$$

respectively.

Level 3: This level is the same as in the descriptive model.

Tak and Morris[51] give the data-dependent posterior propriety conditions of Bayes beta-binomial-logit model for a series of hyper-priors of the hyper-parameters $\boldsymbol{\beta}$ and r. Define a group whose number of success y_i is neither 0 nor n_i as an *interior group* and N_y as the number of interior groups in the total N groups. According to Tak and Morris[51], the full posterior distribution of random effects and hyper-parameters given the Level 3 hyper-prior is proper if and only if there are at least two interior groups in the data and the $N_y \times m$ covariate matrix of the interior groups is of full rank m ($N_y \geq m$). This condition is mild and can be satisfied in most application scenarios.

The likelihood of the hyper-parameters $\boldsymbol{\beta}$ and r is the product of N independent beta-binomial distributions

$$L(\boldsymbol{\beta}, r) = \prod_{i=1}^{N} \binom{n_i}{y_i} \frac{B(y_i + rp_i^E, n_i - y_i + r(1 - p_i^E))}{B(rp_i^E, r(1 - p_i^E))}. \tag{2.91}$$

The posterior distribution of the hyper-parameters given the observed data is

$$p(\boldsymbol{\beta}, r|\boldsymbol{y}) = L(\boldsymbol{\beta}, r)/r^2. \tag{2.92}$$

Before moving on to the next step, they first conduct a transformation of variables on the hyper-parameter r. Let $\alpha = -\log(r)$. The reason for this transformation is that α is more symmetric than r and is distributed on $(-\infty, \infty)$; therefore, α is proper for MLE approximation while r is not. Then, we obtain the transformed posterior distribution of $\boldsymbol{\beta}$ and α:

$$p(\boldsymbol{\beta}, \alpha|\boldsymbol{y}) = L(\boldsymbol{\beta}, r(\alpha))e^{\alpha}. \tag{2.93}$$

To correct the bias when estimating α, a restricted maximum likelihood (REML) type correction is applied by using the Laplace approximation with the Lebesgue measure on $\boldsymbol{\beta}$:

$$\begin{aligned} L(\alpha) &= \int L(\boldsymbol{\beta}, r(\alpha))e^{\alpha} d\boldsymbol{\beta} \\ &= c|\hat{\boldsymbol{H}}_{\alpha}|^{-1/2} e^{\alpha} L(\hat{\boldsymbol{\beta}}_{\alpha}, r(\alpha)), \end{aligned} \tag{2.94}$$

where $\hat{\boldsymbol{H}}_{\alpha}$ is the Hessian Matrix of $\log\left(L\Big(\boldsymbol{\beta}, r(\alpha)\Big)\right)$ evaluated at $\hat{\boldsymbol{\beta}}_{\alpha}$, and $\hat{\boldsymbol{\beta}}_{\alpha}$ is the solution to:

$$\frac{\partial \log(L(\boldsymbol{\beta}, r(\alpha)))}{\partial \boldsymbol{\beta}} = 0 \tag{2.95}$$

at fixed α. To solve equation (2.95), first set an initial value of α. It is easy to get $\hat{\boldsymbol{\beta}}_{\alpha}$ by optimizing the log-likelihood $\log\left(L\Big(\boldsymbol{\beta}, r(\alpha)\Big)\right)$. Inserting $\hat{\boldsymbol{\beta}}_{\alpha}$ into the equation (2.94) yields an updated optimized estimate of α. Repeating this

procedure multiple times will finally yield the estimates of $\hat{\boldsymbol{\beta}}_\alpha$ and $\hat{\alpha}$ and also the Hessian matrix of $\hat{\boldsymbol{\beta}}$ and $\hat{\alpha}$ evaluated at $\hat{\boldsymbol{\beta}}_\alpha$ and $\hat{\alpha}$. It is assumed that the hyper-parameters $\boldsymbol{\beta}$ and α are multivariate normally distributed.

The final goal is to estimate the posterior mean and posterior variance of the random effect p_i

$$E(p_i|\boldsymbol{y}) = E(p_i^*) = (1 - E(B_i|\boldsymbol{y}))\bar{y}_i + E(B_i p_i^E|\boldsymbol{y}), \tag{2.96}$$

and

$$Var(p_i|\boldsymbol{y}) = E(\frac{p_i^*(1-p_i^*)}{r+n_i+1}|\boldsymbol{y}) + Var(p_i^*|\boldsymbol{y}). \tag{2.97}$$

The assumption is that the hyper-parameters $\boldsymbol{\beta}$ and r are independent a posteriori. Thus, both equation (2.96) and equation (2.97) are functions of the posterior moments of B_i and p_i^E.

First use ADM to approximate the posterior distribution of B_i. Since B_i is between 0 and 1, it is appropriate to approximate the posterior distribution of B_i by a beta distribution:

$$B_i|\boldsymbol{y} \sim Beta(a_{i1}, a_{i2}). \tag{2.98}$$

To calculate a_{i1} and a_{i2}, they use a procedure similar to approximating the shrinkage in the gamma-Poisson model. The shrinkage B_i is equal to $e^{-\alpha}/(n_i + e^{-\alpha})$ and the parameter α is approximately normally distributed by the pre-

vious approximation. Then they have

$$\begin{aligned} B_i|\boldsymbol{y} &= \frac{e^{-\alpha}}{n_i + e^{-\alpha}}|\boldsymbol{y} = \frac{e^{-\alpha}/n_i}{1 + e^{-\alpha}/n_i}|\boldsymbol{y} \\ &= \frac{e^{-\alpha-\log(n_i)}}{1 + e^{-\alpha-\log(n_i)}}|\boldsymbol{y} \sim logit - normal(-\hat{\alpha} - \log(n_i), \hat{\sigma}^2_\alpha), \end{aligned} \tag{2.99}$$

where $\hat{\alpha}$ and $\hat{\sigma}^2_\alpha$ can be obtained from the previous approximation.

Assume $B_i|\boldsymbol{y}$ has density $f(B_i|\boldsymbol{y})$ so that the adjusted posterior distribution is $B_i(1-B_i)f(B_i|\boldsymbol{y})$ for beta approximation and define $\mathcal{L}(B_i) = \log\Big(B_i(1-B_i)f(B_i|\boldsymbol{y})\Big)$. Take the first and second derivatives of $\mathcal{L}(B_i)$:

$$\mathcal{L}(B_i) = \log(logit - normal(B_i)B_i(1-B_i)) \tag{2.100}$$

$$= constant - \frac{(logit(B_i) + \hat{\alpha} + log(n_i))^2}{2\hat{\sigma}^2_\alpha}, \tag{2.101}$$

$$\frac{\partial \mathcal{L}(B_i)}{\partial B_i} = \frac{logit(B_i) + \hat{\alpha} + log(n_i)}{\hat{\sigma}^2_\alpha B_i(1-B_i)} = 0 \Rightarrow \hat{B}_i = \frac{e^{-\hat{\alpha}}}{n_i + e^{-\hat{\alpha}}}, \tag{2.102}$$

$$-\ddot{\mathcal{L}}(B_i) = -\left.\frac{\partial^2 \mathcal{L}(B_i)}{\partial B_i^2}\right|_{\hat{B}_i} = \frac{1}{\hat{B}_i^2(1-\hat{B}_i)^2\hat{\sigma}^2_\alpha}. \tag{2.103}$$

Then they apply the ADM approximation procedure. There are two equations for the two unknowns a_{i1} and a_{i2}. Solving the system of equations yields the expressions for the approximation parameters $\hat{a}_{i1}$ and $\hat{a}_{i2}$:

$$\begin{aligned} \frac{\hat{a}_{i1}}{\hat{a}_{i1} + \hat{a}_{i2}} &= \hat{B}_i, \\ \hat{a}_{i1} + \hat{a}_{i2} &= \frac{1}{\hat{B}_i(1-\hat{B}_i)\hat{\sigma}^2_\alpha}. \end{aligned} \tag{2.104}$$

The solution is $\hat{a}_{i1} = \frac{1}{\hat{\sigma}^2_\alpha(1-\hat{B}_i)}$ and $\hat{a}_{i2} = \frac{1}{\hat{\sigma}^2_\alpha \hat{B}_i}$. Now $B_i|\boldsymbol{y}$ can be approximated

by $Beta(\hat{a}_{i1}, \hat{a}_{i2})$. The c-th moment of $B_i|\boldsymbol{y}$ is

$$\hat{E}(B_i^c|\boldsymbol{y}) = \frac{B(\hat{a}_{i1}+c, \hat{a}_{i2})}{B(\hat{a}_{i1}, \hat{a}_{i2})}, \tag{2.105}$$

where B denotes the beta function.

The next step is to approximate the posterior distribution for p_i^E. The unconditional posterior c-th moment of p_i^E is approximated by the conditional posterior moment with $\hat{\alpha}$ substituted for α[23]:

$$E\Big((p_{ik}^E)^c|\boldsymbol{y}\Big) \approx E\Big((p_{ik}^E)^c|\hat{\alpha}, \boldsymbol{y}\Big). \tag{2.106}$$

Use another ADM by assuming the conditional posterior distributions of p_i^E evaluated at $\hat{\alpha}$ are approximately beta distributions:

$$p_i^E|\hat{\alpha}, \boldsymbol{y} \sim Beta(b_{i1}, b_{i2}). \tag{2.107}$$

Then,

$$p_i^E|(\hat{\alpha}, \boldsymbol{y}) = \frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}|(\hat{\alpha}, \boldsymbol{y}) = \frac{G_{i1}}{G_{i1}+G_{i2}}, \tag{2.108}$$

where G_{i1} and G_{i2} are independent random variables following $Gamma(b_{i1}, 1)$ and $Gamma(b_{i2}, 1)$ distributions, respectively. Then,

$$e^{\boldsymbol{x}_i'\boldsymbol{\beta}}|\hat{\alpha}, \boldsymbol{y} \sim \frac{G_{i1}}{G_{i2}}. \tag{2.109}$$

The mean and variance of the ratio of two independent Gamma distributions are

$$E(e^{\boldsymbol{x}_i'\boldsymbol{\beta}}|\hat{\alpha},\boldsymbol{y}) = E\left(\frac{G_{i1}}{G_{i2}}\right) = \frac{b_{i1}}{b_{i2}-1} = \eta_i, \tag{2.110}$$

$$Var(e^{\boldsymbol{x}_i'\boldsymbol{\beta}}|\hat{\alpha},\boldsymbol{y}) = Var\left(\frac{G_{i1}}{G_{i2}}\right) = \frac{\eta_i(1+\eta_i)}{b_{i2}-2}. \tag{2.111}$$

From the previous approximation, $\boldsymbol{\beta}$ is multivariate normally distributed with mean $\hat{\boldsymbol{\beta}}$ and covariance matrix $\hat{\Sigma}$. Then, the mean and variance of $log-normal$ distributions are easy to compute:

$$\hat{E}(e^{\boldsymbol{x}_i'\boldsymbol{\beta}}|\hat{\alpha},\boldsymbol{y}) = \exp(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}} + \boldsymbol{x}_i'\hat{\Sigma}\boldsymbol{x}_i/2) = \hat{\eta}_i, \tag{2.112}$$

$$\widehat{Var}(e^{\boldsymbol{x}_i'\boldsymbol{\beta}}|\hat{\alpha},\boldsymbol{y}) = \hat{\eta}_i^2(e^{\boldsymbol{x}_i'\hat{\Sigma}\boldsymbol{x}_i} - 1). \tag{2.113}$$

By matching the means in equation (2.110) and equation (2.112) and the variances in equation (2.111) and equation (2.113), there are two equations for two unknown b_{i1} and b_{i2}

$$E(e^{\boldsymbol{x}_i'\boldsymbol{\beta}_j}|\hat{\alpha},\boldsymbol{y}) = \hat{E}(e^{\boldsymbol{x}_i'\boldsymbol{\beta}_j}|\hat{\alpha},\boldsymbol{y}) \tag{2.114}$$

and

$$Var(e^{\boldsymbol{x}_i'\boldsymbol{\beta}}|\hat{\alpha},\boldsymbol{y}) = \widehat{Var}(e^{\boldsymbol{x}_i'\boldsymbol{\beta}}|\hat{\alpha},\boldsymbol{y}). \tag{2.115}$$

The solution for b_{i1} and b_{i2} are

$$\hat{b}_{i2} = \frac{1+\hat{\eta}_i}{\hat{\eta}_i(e^{\boldsymbol{x}_i'\hat{\Sigma}\boldsymbol{x}_i}-1)} + 2, \tag{2.116}$$

and

$$\hat{b}_{i1} = \hat{\eta}_i(\hat{b}_{i2}-1). \tag{2.117}$$

Each $p_i^E|\hat{\alpha}, \boldsymbol{y}$ has approximately a $beta(\hat{b}_{i1}, \hat{b}_{i2})$ distribution. Then,

$$\hat{E}\Big((p_i^E)^c|\hat{\alpha}, \boldsymbol{y}\Big) = \frac{B(\hat{b}_{i1}+c, \hat{b}_{i2})}{B(\hat{b}_{i1}, \hat{b}_{i2})},\ c \geq 0. \tag{2.118}$$

Now move back to the estimation of posterior mean and posterior variance of p_i. The assumption is that $\boldsymbol{\beta}$ and α are independent a posteriori; therefore

$$E(p_i|\boldsymbol{y}) = (1-E(B_i|\boldsymbol{y}))\bar{y}_i + E(B_i|\boldsymbol{y})E(p_i^E|\boldsymbol{y}), \tag{2.119}$$

$$\begin{aligned}
Var(p_i|\boldsymbol{y}) &= E\Big(\frac{p_i^*(1-p_i^*)}{n_i+r+1}|\boldsymbol{y}\Big) + Var\Big(p_i^*|\boldsymbol{y}\Big) \\
&= E\Big(\frac{p_i^*(1-p_i^*)}{n_i+r+1}|\boldsymbol{y}\Big) + Var\Big(B_i(\bar{y}_i - p_i^E)|\boldsymbol{y}\Big) \\
&\approx E\Big(\frac{p_i^*(1-p_i^*)(1-B_i)}{n_i}|\boldsymbol{y}\Big) + Var\Big(B_i(\bar{y}_i - p_i^E)|\boldsymbol{y}\Big) \\
&= \Big\{(1-\bar{y}_i)\bar{y}_i[1-E(B_i|\boldsymbol{y})] \\
&\quad + (2\bar{y}_i - 1)E(B_i(1-B_i)|\boldsymbol{y})(\bar{y}_i - E(p_i^E|\boldsymbol{y})) \\
&\quad + E(B_i^2(1-B_i)|\boldsymbol{y})E((\bar{y}_i - p_i^E)^2|\boldsymbol{y})\Big\}/n_i + Var(B_i(\bar{y}_i - p_i^E)|\boldsymbol{y}).
\end{aligned} \tag{2.120}$$

The approximation in equation (2.120) is a first-order Taylor approximation. By inserting the shrinkage and expected probability moment estimates as in equation (2.105) and equation (2.118) into equation (2.119) and equation (2.120), the estimated posterior mean and posterior variance of the random binomial probabilities can be computed.

Denote $\hat{\mu}_{p_i} = \hat{E}(p_i|\boldsymbol{y})$ and $\hat{\sigma}^2_{p_i} = \widehat{Var}(p_i|\boldsymbol{y})$ and assume $p_i|\boldsymbol{y}$ is approximately $beta(t_{i1}, t_{i2})$ distributed. The estimates of t_{i1} and t_{i2} are as follows:

$$\hat{t}_{i1} = \Big(\frac{\hat{\mu}_{p_i}(1-\hat{\mu}_{p_i})}{\hat{\sigma}^2_{p_i}} - 1\Big)\hat{\mu}_{p_i}, \tag{2.121}$$

and

$$\hat{t}_{i2} = \Big(\frac{\hat{\mu}_{p_i}(1-\hat{\mu}_{p_i})}{\hat{\sigma}^2_{p_i}} - 1\Big)(1-\hat{\mu}_{p_i}). \tag{2.122}$$

Finally, the assumed unconditional posterior distribution of random effect for the beta-binomial-logit model is

$$p_i|\boldsymbol{y} \sim Beta(\hat{t}_{i1}, \hat{t}_{i2}) \tag{2.123}$$

2.5 Discussion

The ADM provides an alternative method for parameter estimation in hierarchical Bayes models to MCMC and other procedures (e.g., EB-MLE or EB-REML). It has been applied to Poisson-gamma, normal-normal, and binomial-

beta models as discussed in Section 2.2 to Section 2.4 and has been proven to have some attractive advantages over a range of parameter estimation procedures[13;38]. The advantage of the ADM compared to MCMC is firstly the overwhelmingly fast speed while it still maintains the accuracy of the estimates as observed in empirical studies. Another advantage over MCMC is that the ADM generates the same result each time a model is applied to the same dataset which MCMC does not do. And because the ADM adopts appropriate adjustments to the likelihood function and includes the ADM approximation when estimating the first level parameters, it has better operating characteristics (e.g., coverage rate, interval width and squared error risks), as proven by previous studies[13], compared to EB procedures. Multiplying the likelihood by the third level hyper-prior of the hyper-parameters prevents posterior impropriety, which can happen in the EB procedures when estimating the posterior distribution of the hyper-parameters. Also, the ADM considers the variance in the estimates of the hyper-parameters when estimating the first level parameters, which EB-plugin procedures do not do. Thus, the ADM generates wider intervals for the first level parameters, which partly explains the higher coverage rate of the ADM compared to EB procedures. Because of all the above favorable characteristics of the ADM, it is desirable to extend the ADM to more hierarchical Bayes models with different distributions. We have conducted a research on an ADM to estimate the multinomial probabilities in the multinomial-Dirichlet-logit model in Chapter 3 of this dissertation.

Chapter 3: Multinomial-Dirichlet-Logit Model

3.1 Introduction

This chapter proposes a procedure for multinomial data analysis. We start with a description of the multinomial-Dirichlet-logit model in two mathematically equivalent forms[13;24] and obtain closed-forms of the approximating posterior distributions of the multinomial probabilities. Section 3.2 and Section 3.3 present the multinomial-Dirichlet-logit model in descriptive and inferential forms, respectively. These two forms are mathematically equivalent in the sense that they generate the same joint distribution of the data and the first level parameters conditional on the hyper-parameters. Section 3.4 explains the selection of the hyper-prior distribution for the hyper-parameters, and provides a sufficient condition on the data for the posterior distribution of the hyper-parameters to be proper. Through careful adjustments to the likelihood function, Section 3.5 approximates the joint distribution of the transformed hyper-parameters by a multivariate normal distribution. Section 3.6 applies the ADM to approximate the posterior distributions of the multinomial probabilities. Section 3.7 briefly concludes this chapter.

3.2 The Descriptive Model

The observed data is a set of vectors $\{\boldsymbol{y_i}\}$, $i = 1, \ldots, N$, where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iK})^T$ and y_{ik} is the non-negative integer count for the k-th category, $k = 1, \ldots, K$ with $K \geq 3$. The condition that y_{ik} is an interger can be relaxed for the ADM. An example of handling non-integer counts in a multinomial distribution with our proposed approach can be found in Chapter 5. There are three levels in the multinomial-Dirichlet-logit model. Level 1 specifies the multinomial distribution for $\boldsymbol{y}_i$ given the multinomial probabilities $\boldsymbol{p}_i = (p_{i1}, \ldots, p_{iK})$, $i = 1, \ldots, N$. The main goal of this dissertation is to make inferences about the multinomial probabilities $\boldsymbol{p}_i = (p_{i1}, \ldots, p_{iK})$. Level 2 assigns the Dirichlet distribution to $\boldsymbol{p}_i$ given the hyper-parameters $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{K-1})$ and $r \in \mathbb{R}^+$, where $\boldsymbol{\beta} \in \mathbb{R}^{(K-1)\times q}$ is the set of regression coefficient vectors for the first (K-1) categories and $\boldsymbol{\beta}_k \in \mathbb{R}^q$ for $k = 1, \ldots, K-1$ with q being the number of covariates in the regression. Level 3 states the hyper-prior distribution for the hyper-parameters $\boldsymbol{\beta}$ and r.

Level 1: The observations $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iK})$, $i = 1, \ldots, N$, have independent multinomial distributions given the individual parameters $\boldsymbol{p}_i = (p_{i1}, \ldots, p_{iK})$, $i = 1, \ldots, N$.

$$(y_{i1}, \ldots, y_{iK}) | p_{i1}, \ldots, p_{iK} \sim multinomial(n_i, p_{i1}, \ldots, p_{iK}), \qquad (3.1)$$

where $n_i = \sum_{k=1}^{K} y_{ik}$ and $\sum_{k=1}^{K} p_{ik} = 1$.

Level 2: The multinomial probabilities $\boldsymbol{p}_i = (p_{i1}, \ldots, p_{iK})$ follow conjugate Dirichlet distributions for $i = 1, \ldots, N$ independently, given the hyper-parameters $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{K-1})$ and r:

$$(p_{i1}, \ldots, p_{iK}) | \boldsymbol{\beta}, r \sim Dirichlet(rp_{i1}^E, \ldots, rp_{iK}^E), \tag{3.2}$$

where the synthetic probabilities $\boldsymbol{p}_i^E = (p_{i1}^E, \ldots, p_{iK}^E)$ are given by

$$p_{ik}^E = \frac{e^{\boldsymbol{x}_i' \boldsymbol{\beta}_k}}{1 + \sum_{j=1}^{K-1} e^{\boldsymbol{x}_i' \boldsymbol{\beta}_j}}, \tag{3.3}$$

for $k = 1, \ldots, K-1$ and

$$p_{iK}^E = \frac{1}{1 + \sum_{j=1}^{K-1} e^{\boldsymbol{x}_i' \boldsymbol{\beta}_j}}, \tag{3.4}$$

so that $\boldsymbol{p}_i^E$ satisfies the condition that $\sum_{k=1}^{K} p_{ik}^E = 1$ and $\boldsymbol{\beta}_k \in \mathbb{R}^q$ is the regression coefficient vector for categories $k = 1, \ldots, K-1$. Here $\boldsymbol{x}_i \in \mathbb{R}^q$ is the vector of known covariates for individual group i, for $i = 1, \ldots, N$. The hyper-parameter r accounts for between-individual variability. Following the terminology Christiansen and Morris (1997)[13], we call r the variance component.

Level 3: For the hyper-prior, we assume a flat distribution for the hyper-

parameter $\boldsymbol{\beta}$ and a flat distribution for the reciprocal of r:

$$\boldsymbol{\beta}_k \sim Uniform\ on\ \mathbb{R}^q, \tag{3.5}$$

for $k = 1, \ldots, K-1$, and

$$1/r \sim Uniform(0, \infty). \tag{3.6}$$

The choice of the improper hyper-prior distribution for $\boldsymbol{\beta}$ is standard. The selection of the hyper-prior distribution for r is to correct the estimation problem that the posterior mode (MLE in the classical terminology) of r can occur at infinity. When this happens, the ADM approximation to the posterior mean of the shrinkage $B_i = r/(r + n_i)$ occurs at the boundary point, which will affect the accuracy of the ADM approximation. This hyper-prior will eliminate the posterior impropriety under a mild condition on the data. This will be discussed in detail in Section 3.4.

3.3 The Inferential Model

The marginal distributions of the observed data $\{\boldsymbol{y}_i\}$, $i = 1, \ldots, N$, given the hyper-parameters $\boldsymbol{\beta}$ and r, are independent Dirichlet-multinomial distributions. They are derived by integrating the first level parameter $\boldsymbol{p}_i$ out and

the density is given by:

$$f(\boldsymbol{y}_i|\boldsymbol{\beta}, r) = \frac{(n_i!)\Gamma(r)}{\Gamma(n_i + r)} \prod_{k=1}^{K} \frac{\Gamma(y_{ik} + rp_{ik}^E)}{(y_{ik}!)\Gamma(rp_{ik}^E)}, \tag{3.7}$$

where the notation $\Gamma(x)$ stands for the gamma function.

Level 1: The observations $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iK})$, given the hyper-parameters $\boldsymbol{\beta}$ and r, have independent Dirichlet-multinomial distributions for $i = 1, \ldots, N$ with densities as in equation (3.7),

$$(y_{i1}, \ldots, y_{iK})|\boldsymbol{\beta}, r \sim DM(n_i, rp_{i1}^E, \ldots, rp_{iK}^E). \tag{3.8}$$

For notational simplicity, we use DM to stand for the Dirichlet-multinomial distribution.

Level 2: Because of conjugacy, the conditional posterior distributions for the multinomial probability parameters $\boldsymbol{p}_i = (p_{i1}, \ldots, p_{iK})$, given the hyper-parameters $\boldsymbol{\beta}$ and r and the data $\boldsymbol{y}$, are independent Dirichlet distributions with updated parameters:

$$(p_{i1}, \ldots, p_{iK})|\boldsymbol{\beta}, r, \boldsymbol{y} \sim Dirichlet(n_i\bar{y}_{i1} + rp_{i1}^E, \ldots, n_i\bar{y}_{iK} + rp_{iK}^E), \tag{3.9}$$

where $\bar{y}_{ik} = y_{ik}/n_i$ is the observed proportion of category k in group i, $i = 1, \ldots, N$ and $k = 1, \ldots, K$. The means, variances and covariances of condi-

tional posterior distribution for $\boldsymbol{p}_i = (p_{i1}, \ldots, p_{iK})$ are given by:

$$p^*_{ik} = E(p_{ik}|\boldsymbol{\beta}, r, \boldsymbol{y}) = (1 - B_i)\bar{y}_{ik} + B_i p^E_{ik}, \tag{3.10}$$

$$Var(p_{ik}|\boldsymbol{\beta}, r, \boldsymbol{y}) = \frac{p^*_{ik}(1 - p^*_{ik})}{n_i + r + 1}, \tag{3.11}$$

$$Cov(p_{il}, p_{im}|\boldsymbol{\beta}, r, \boldsymbol{y}) = -\frac{p^*_{il}p^*_{im}}{n_i + r + 1} \; with \; l \neq m, \tag{3.12}$$

where $B_i = r/(r + n_i)$, known as the shrinkage factor for group i. The hyper-parameter $r > 0$ can also be explained as the unobserved total hyper-prior counts for group i.

Level 3 remains the same as in the descriptive model.

3.4 Posterior Propriety

The joint posterior density $f(\boldsymbol{\beta}, r|\boldsymbol{y})$ of the hyper-parameters $\boldsymbol{\beta}$ and r is given by

$$f(\boldsymbol{\beta}, r|\boldsymbol{y}) \propto L(\boldsymbol{\beta}, r)/r^2, \tag{3.13}$$

where the likelihood function is the product of the N independent Dirichlet-multinomial densities

$$L(\boldsymbol{\beta}, r) = \prod_{i=1}^{N} \left(\frac{(n_i!)\Gamma(r)}{\Gamma(n_i + r)} \prod_{k=1}^{K} \frac{\Gamma(y_{ik} + rp^E_{ik})}{(y_{ik}!)\Gamma(rp^E_{ik})} \right). \tag{3.14}$$

The propriety of the posterior is data-dependent. In this section, we provide a

sufficient condition on the data for the posterior to be proper. This sufficient condition is mild and can be satisfied in many application scenarios. The lemmas and the theorems in this sections are all for integer counts y_{ik}. Some extensions to non-integer counts are proved in Appendix A.5.

Definition 3.1 *Let d_i, $1 \leq d_i \leq K (K \geq 3)$, denote the number of non-zeros in group i of the data. There can be three types of groups in the data: (1) interior group ($d_i = K$) with $y_{ik} \geq 1$ for all $k = 1, \ldots, K$; (2) intermediate group ($2 \leq d_i \leq K-1$) with at least one zero and at least two non-zeros in the group; and (3) extreme group ($d_i = 1$) with all the mass n_i in group i concentrating within one category. Let the symbol W_i denote the set of indices of the categories with positive counts in group i, $W_i \subseteq \{1, \ldots, K\}$, and let d_i be the cardinality of the set W_i.*

Definition 3.2 *The symbol $W_y \subseteq \{1, \ldots, N\}$ denotes the set of indices corresponding to interior groups in the data and N_y denotes the number of interior groups, that is, the length of the set W_y. Use the symbol W_y^c to denote the set of indices of intermediate and extreme groups and let $(N - N_y)$ be the number of intermediate and extreme groups in the data. The notation $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)^T$ refers to the $N \times q$ covariate matrix of all groups ($N \geq q$) and $\boldsymbol{X}_y$ is the $N_y \times q$ covariate matrix of the interior groups.*

Lemma 3.1 *The lower and upper bounds for the Dirichlet-multinomial probability mass function for interior group i with respect to $\boldsymbol{\beta}$ and r are given*

by

$$\frac{r^{n_i}}{(n_{max}+r)^{n_i}} \prod_{k=1}^{K} (p_{ik}^E)^{y_{ik}} \tag{3.15}$$

and

$$\frac{r^2}{(r+1)(r+2)} \prod_{k=1}^{K} p_{ik}^E, \tag{3.16}$$

respectively, up to a constant multiple, where $n_{max} = \max\{n_1, \ldots, n_N\}$. *Those for intermediate group* i *are given by*

$$\frac{r^{n_i}}{(n_{max}+r)^{n_i}} \prod_{k \in W_i} (p_{ik}^E)^{y_{ik}} \tag{3.17}$$

and

$$\frac{r}{r+1} \prod_{k \in W_i} p_{ik}^E, \tag{3.18}$$

respectively, up to a constant multiple. Those for extreme group i *are* $(p_{ij}^E)^{n_i}$ *and* p_{ij}^E, *respectively when* j *is the index of the category with* $y_{ij} = n_i$, *up to a constant multiple.*

Proof. See Appendix A.1.

As can be seen from Lemma 3.1, both the lower and upper bounds of Dirichlet-multinomial probability mass function can be factored into a function of $\boldsymbol{\beta}$ and a function of r. The likelihood $L(r, \boldsymbol{\beta})$ is the product of the individual Dirichlet-multinomial probability mass functions; thus, the lower and upper bounds of $L(r, \boldsymbol{\beta})$ can also be factored into functions of $\boldsymbol{\beta}$ and r, respectively.

Lemma 3.2 *When all groups are interior, $L(r, \boldsymbol{\beta})$ can be bounded from below and above by*

$$c\Big(\frac{r}{n_{max}+r}\Big)^{\sum_{i=1}^{N} n_i} \prod_{i=1}^{N}\prod_{k=1}^{K}(p_{ik}^{E})^{y_{ik}} \leq L(r,\boldsymbol{\beta}) \leq c\frac{r^{2N}}{(r+1)^{N}(r+2)^{N}} \prod_{i=1}^{N}\prod_{k=1}^{K} p_{ik}^{E}, \tag{3.19}$$

where c is a constant that does not depend on $\boldsymbol{\beta}$ and r.

Proof. See Appendix A.2.

Now we can show that with a flat prior on r, the posterior, which is proportional to the likelihood, is improper. The integral of the part with respect to r in the lower bound of equation (3.19),

$$\int_0^{\infty} \Big(\frac{r}{n_{max}+r}\Big)^{\sum_{i=1}^{N} n_i} dr, \tag{3.20}$$

does not converge. Thus, the posterior mode (MLE in the classical terminology) of r can occur at infinity when the sample size is small. This explains the selection of a hyper-prior on r other than a flat function. Theorem 3.1 below proves a sufficient condition for posterior propriety for a dataset with all interior groups.

Theorem 3.1 *When all groups are interior in the data, the posterior density function of hyper-parameters $p(\boldsymbol{\beta}, r|\boldsymbol{y})$, equipped with $f(r) \propto 1/r^2$ and independently an improper flat hyper-prior density on $\boldsymbol{\beta}$, $g(\boldsymbol{\beta}) \propto 1$, is proper if the covariate matrix $\boldsymbol{X}$ is of full rank q.*

Proof. See Appendix A.3.

Corollary 3.1 *When interior groups co-exist with intermediate and/or extreme groups, the propriety of the posterior can be determined solely by the interior groups.*

Proof. See Appendix A.4.

Thus the posterior density $f(\boldsymbol{\beta}, r|\boldsymbol{y})$ is proper provided that there is at least one interior group (e.g., all y_{ik}'s, $k = 1, \ldots, K$, are greater than or equal to 1) in the data and that the $N_y \times q$ sub-matrix $\boldsymbol{X}_y$ of $\boldsymbol{X}$ is of full rank.

We also prove the bounds for the Dirichlet-multinomial probability mass function for interior and intermediate groups with non-integer counts in Appendix A.5. The rest of the proof for the sufficient condition for the non-integer counts are very similar to the case of integer counts and will not be given in detail.

3.5 Distribution of Hyper-parameters

The likelihood function of the hyper-parameters $\boldsymbol{\beta}$ and r is the product of the N DM densities as in equation (3.7):

$$L(\boldsymbol{\beta}, r) = \prod_{i=1}^{N} \left(\frac{(n_i!)\Gamma(r)}{\Gamma(n_i + r)} \prod_{k=1}^{K} \frac{\Gamma(y_{ik} + r p_{ik}^E)}{(y_{ik}!)\Gamma(r p_{ik}^E)} \right). \tag{3.21}$$

The log-likelihood function of $L(\boldsymbol{\beta}, r)$ is given by

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}, r) &= C + \sum_{i=1}^{N} \Bigg(\log\Big(\Gamma(r)\Big) - \log\Big(\Gamma(n_i + r)\Big) \\ &\qquad + \sum_{k=1}^{K} \Bigg(\log\Big(\Gamma(y_{ik} + rp_{ik}^E)\Big) - \log\Big(\Gamma(rp_{ik}^E)\Big) \Bigg) \Bigg) \\ &= C + N \log\Big(\Gamma(r)\Big) - \sum_{i=1}^{N} \log\Big(\Gamma(n_i + r)\Big) \\ &\quad + \sum_{i=1}^{N} \sum_{k=1}^{K} \Bigg(\log\Big(\Gamma(y_{ik} + rp_{ik}^E)\Big) - \log\Big(\Gamma(rp_{ik}^E)\Big) \Bigg),\end{aligned} \tag{3.22}$$

where C is a constant independent of $\boldsymbol{\beta}$ and r.

Fix $r > 0$ and define $\alpha = -\log(r)$ throughout, because the distribution of α is more symmetric than r and α is defined on the real line without any boundary issues. Normal approximation would be more accurate for α than for r. The log-likelihood of $\boldsymbol{\beta}$ alone, with r fixed, involves only the last term of (3.22). Then the score function, the q-dimensional gradient of log-likelihood (3.22) with respect to the regression coefficient vector $\boldsymbol{\beta}_k$ for $k = 1, \ldots, K-1$, is given by

$$\frac{\partial \mathcal{L}(\boldsymbol{\beta}, r)}{\partial \boldsymbol{\beta}_k} = \sum_{i=1}^{N} rp_{ik}^E \Big\{ \Big[\psi(y_{ik}+rp_{ik}^E) - \psi(rp_{ik}^E)\Big] - \sum_{j=1}^{K} p_{ij}^E \Big[\psi(y_{ij}+rp_{ij}^E) - \psi(rp_{ij}^E)\Big] \Big\} \boldsymbol{x}_i^T, \tag{3.23}$$

where $\psi(x) = (d/dx) \log \Gamma(x)$ is the digamma function. The second derivatives of the log-likelihood function (3.22) with respect to $\boldsymbol{\beta}_k$, $k = 1, \ldots, K-1$, is

given by

$$\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta}, r)}{\partial(\boldsymbol{\beta}_k \boldsymbol{\beta}_k')} = \sum_{i=1}^{N} \Big\{ r^2 (p_{ik}^E)^2 (1 - 2p_{ik}^E) \Big[\psi_1(y_{ik} + r p_{ik}^E) - \psi_1(r p_{ik}^E) \Big] + r p_{ik}^E (1 - 2p_{ik}^E) \Big[\psi(y_{ik} + r p_{ik}^E) - \psi(r p_{ik}^E) \Big] + r^2 (p_{ik}^E)^2 \sum_{j=1}^{K} (p_{ij}^E)^2 \Big[\psi_1(y_{ij} + r p_{ij}^E) - \psi_1(r p_{ij}^E) \Big] - r p_{ik}^E (1 - 2p_{ik}^E) \sum_{j=1}^{K} p_{ij}^E \Big[\psi(y_{ij} + r p_{ij}^E) - \psi(r p_{ij}^E) \Big] \Big\} \boldsymbol{x}_i \boldsymbol{x}_i^T, \tag{3.24}$$

where $\psi_1(x) = (d^2/dx^2) \log \Gamma(x)$ is the trigamma function. Writing the expression enclosed by the curly brackets in equation (3.24) as a_{ik}, equation (3.24) can be written in matrix form:

$$\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta}, r)}{\partial(\boldsymbol{\beta}_k \boldsymbol{\beta}_k')} = \boldsymbol{X'} \boldsymbol{D}_k \boldsymbol{X}, \ k = 1, \ldots, K-1, \tag{3.25}$$

where $\boldsymbol{D}_k$ is the $N \times N$ diagonal matrix with the i^{th} diagonal element being a_{ik}, $i = 1, \ldots, N$, and $\boldsymbol{X}$ is the $N \times q$ matrix of covariates. The mixed second derivative with respect to the regression coefficients for the l^{th} and the m^{th}

categories, $\boldsymbol{\beta}_l$ and $\boldsymbol{\beta}_m$, is given by:

$$\begin{aligned}\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta}, r)}{\partial(\boldsymbol{\beta}_l \boldsymbol{\beta}'_m)} = \sum_{i=1}^{N} \Big\{ & - r^2 (p_{il}^E)^2 p_{im}^E \Big[\psi_1(y_{il} + r p_{il}^E) - \psi_1(r p_{il}^E)\Big] \\ & - r p_{il}^E p_{im}^E \Big[\psi(y_{il} + r p_{il}^E) - \psi(r p_{il}^E)\Big] \\ & - r^2 p_{il}^E (p_{im}^E)^2 [\psi_1(y_{im} + r p_{im}^E) - \psi_1(r p_{im}^E)] \\ & - r p_{il}^E p_{im}^E [\psi(y_{im} + r p_{im}^E) - \psi(r p_{im}^E)] \\ & + r^2 p_{il}^E p_{im}^E \sum_{j=1}^{K} (p_{ij}^E)^2 \Big[\psi_1(y_{ij} + r p_{ij}^E) - \psi_1(r p_{ij}^E)\Big] \\ & + 2 r p_{il}^E p_{im}^E \sum_{j=1}^{K} p_{ij}^E \Big[\psi(y_{ij} + r p_{ij}^E) - \psi(r p_{ij}^E)\Big] \Big\} \boldsymbol{x}_i \boldsymbol{x}_i^T, \end{aligned} \tag{3.26}$$

where $l \neq m$. Similar to the case for the second derivatives with respect to $\boldsymbol{\beta}_k$, the mixed second derivative (3.26) can also be written in matrix form. Define $\boldsymbol{D}_{lm}$ as the diagonal matrix with the i^{th} diagonal term being a_{ilm}, $i = 1, \ldots, N$, where a_{ilm} is equal to the expression enclosed by the curly brackets in equation (3.26); then

$$\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta}, r)}{\partial \boldsymbol{\beta}_l \partial \boldsymbol{\beta}'_m} = \boldsymbol{X}' \boldsymbol{D}_{lm} \boldsymbol{X}. \tag{3.27}$$

From above, the second derivative of the log-likelihood (3.22) with respect to $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{K-1})$ is $\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta}, r)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}$, which has second derivatives (3.24) for $k = 1, \ldots, K-1$ as the blocks in the diagonal. The closed-form first and second derivatives with respect to r are complicated and are not necessary since the optimal hyper-parameter estimates are computed using the *optim*() function in R. The second derivative of the log-likelihood (3.22) with respect to $\boldsymbol{\beta}$ is given because it is used as an adjustment to the likelihood function (3.21) in

order to increase the computation speed without sacrificing the accuracy of the estimates.

With the transformation $\alpha = -\log(r)$, the posterior density of $\boldsymbol{\beta}$ and α is given by

$$f(\boldsymbol{\beta}, \alpha|\boldsymbol{y}) \propto e^{\alpha} L(\boldsymbol{\beta}, r(\alpha)). \tag{3.28}$$

In Section 3.2.3, we prove that this posterior is proper with the adjustment of the Level 3 hyper-prior (3.5). Moreover, adjustment ensures that the mode occurs at a finite value for α. The second adjustment is the restricted maximum likelihood (REML) type correction by Laplace approximation with a Lebesgue measure on $\boldsymbol{\beta}$,

$$f_2(\boldsymbol{\beta}, \alpha|\boldsymbol{y}) = \int f(\boldsymbol{\beta}, \alpha|\boldsymbol{y}) d\boldsymbol{\beta} \quad = c|\boldsymbol{H}|^{-1/2} e^{\alpha} L(\boldsymbol{\beta}, r(\alpha)), \tag{3.29}$$

where

$$\boldsymbol{H} = -\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta}, r(\alpha))}{\partial(\boldsymbol{\beta}\boldsymbol{\beta}')}. \tag{3.30}$$

The REML type adjustment is used to correct for the bias in the posterior mode (MLE in the classical terminology) of α, which results from ignoring the loss of degrees of freedom when estimating $\boldsymbol{\beta}$. This problem is severe with small sample size N and large number of regressors q. This REML type adjustment has been proved to produce a different and better estimate for α by previous studies[19;28]. Because this REML correction $|\boldsymbol{H}|^{-1/2}$ is complex and only approximate, we introduce the third adjustment. The geometric mean

of the $\boldsymbol{H}$ eigenvalues, $|\boldsymbol{H}|^{1/(q(K-1))}$, is approximated by a constant multiple of the geometric mean of the $N(K-1)$ values of a_{ik}. Thus, the approximate logarithm of the adjusted posterior (3.29) is given by

$$\mathcal{L}_R(\boldsymbol{\beta}, r(\alpha)) = c_2 + \alpha + \mathcal{L}(\boldsymbol{\beta}, r(\alpha)) - \frac{q}{2N}\sum_{i=1}^{N}\sum_{k=1}^{K-1} log(|a_{ik}|), \tag{3.31}$$

where c_2 is a constant independent of $\boldsymbol{\beta}$ and α. Then, the next analysis is similar to that for maximum likelihood. The distribution of the hyper-parameters $\boldsymbol{\beta}$ and $\alpha = -log(r)$ can be approximated by a joint multivariate normal:

$$\begin{pmatrix}\boldsymbol{\beta}\\ \alpha\end{pmatrix} |data \sim N_{(K-1)\times q+1}\left[\hat{\boldsymbol{\mu}} = \begin{pmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\alpha}\end{pmatrix}, \hat{\boldsymbol{\Sigma}}\right], \tag{3.32}$$

where $\hat{\boldsymbol{\mu}}$ optimizes $\mathcal{L}_R(\boldsymbol{\beta}, r(\alpha))$ and $\hat{\boldsymbol{\Sigma}}$ is the inverse of the Hessian matrix of $-\mathcal{L}_R(\boldsymbol{\beta}, r(\alpha))$ at $\hat{\boldsymbol{\mu}}$. Both $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ can be computed easily using the R function $optim()$.

3.6 Distributions of the Multinomial Probabilities

The parameters of interest are the multinomial probabilities $\boldsymbol{p}_i = (p_{i1}, \ldots, p_{iK})$ in the multinomial distribution. Thus the ultimate goal is to approximate the posterior distribution of $(p_{i1}, \ldots, p_{iK})|\boldsymbol{y}$ and approximate the posterior means $E(p_{ik}|\boldsymbol{y})$ and the posterior variances $Var(p_{ik}|\boldsymbol{y})$ of the multinomial probabilities. From the conditional posterior mean (3.10) and conditional

posterior variance (3.11), we have

$$\begin{aligned} E(p_{ik}|\boldsymbol{y}) &= E_{\boldsymbol{\beta},\alpha|\boldsymbol{y}}\Big(E(p_{ik}|\boldsymbol{\beta},\alpha,\boldsymbol{y})\Big) = E_{\boldsymbol{\beta},\alpha|\boldsymbol{y}}(p^*_{ik}) \\ &= E_{\boldsymbol{\beta},\alpha|\boldsymbol{y}}\Big((1-B_i)\bar{y}_{ik} + B_i p^E_{ik}\Big) \end{aligned} \tag{3.33}$$

and

$$\begin{aligned} Var(p_{ik}|\boldsymbol{y}) &= E_{\boldsymbol{\beta},\alpha|\boldsymbol{y}}\Big(Var(p_{ik}|\boldsymbol{\beta},\alpha,\boldsymbol{y})\Big) + Var_{\boldsymbol{\beta},\alpha|\boldsymbol{y}}\Big(E(p_{ik}|\boldsymbol{\beta},\alpha,\boldsymbol{y})\Big) \\ &= E_{\boldsymbol{\beta},\alpha|\boldsymbol{y}}\Big(\frac{p^*_{ik}(1-p^*_{ik})}{n_i+r+1}\Big) + Var_{\boldsymbol{\beta},\alpha|\boldsymbol{y}}\Big(p^*_{ik}\Big). \end{aligned} \tag{3.34}$$

We assume that the hyper-parameters $\boldsymbol{\beta}$ and α are independent *a posteriori*. Under this assumption, both the posterior mean (3.33) and the posterior variance (3.34) are functions of the moments $E(B_i^c|\boldsymbol{y})$ and $E\Big((p^E_{ik})^c|\boldsymbol{y})\Big)$, c being a positive integer. Therefore, we want to approximate the distributions of $B_i|\boldsymbol{y}$ and $p^E_{ik}|\boldsymbol{y}$.

3.6.1 Posterior Distribution of Shrinkage B_i

Since the support of $B_i|\boldsymbol{y}$ is between 0 and 1, it is not appropriate to approximate the distribution of $B_i|\boldsymbol{y}$ by a normal distribution and it is more reasonable to approximate it by a beta distribution. In this case, ADM approximation provides the solution. The density $f(B_i|\boldsymbol{y})$

is approximately a logit-normal distribution

$$\begin{aligned} B_i &= \frac{r}{r+n_i} = \frac{e^{-\alpha}}{e^{-\alpha}+n_i} = \frac{e^{-\alpha}}{e^{-\alpha}+e^{log(n_i)}} = \frac{e^{-\alpha-log(n_i)}}{1+e^{-\alpha-log(n_i)}} \\ &\sim Logit-Normal(-\hat{\alpha}-log(n_i),\ \hat{\sigma}_\alpha^2). \end{aligned} \tag{3.35}$$

The last step in equation (3.35) is because α is distributed with $normal(\hat{\alpha}, \hat{\sigma}_\alpha^2)$ from distribution (3.32). There is no analytical form for the moment of a logit-normal distribution. We propose to approximate the logit-normal distribution $f(B_i|\boldsymbol{y})$ by the *beta* distribution with parameters a_{i1} and a_{i2}:

$$B_i|\boldsymbol{y} \sim Beta(a_{i1}, a_{i2}). \tag{3.36}$$

We will follow the steps for ADM approximation to obtain the estimates for a_{i1} and a_{i2}. The adjusted density is $B_i(1-B_i)f(B_i|\boldsymbol{y})$. This is the density to be approximated, $f(B_i|\boldsymbol{y})$, multiplied by the binomial distribution adjustment factor function $B_i(1-B_i)$. Now we define $\mathcal{L}(B_i) = \log\Big(B_i(1-B_i)f(B_i|\boldsymbol{y})\Big)$. Considering the first and second derivatives of $\mathcal{L}(B_i)$

$$\begin{aligned} \mathcal{L}(B_i) &= \log(logit-normal(B_i)B_i(1-B_i)) \\ &= constant - \frac{(logit(B_i)+\hat{\alpha}+log(n_i))^2}{2\hat{\sigma}_\alpha^2}; \end{aligned} \tag{3.37}$$

we obtain

$$\frac{\partial \mathcal{L}(B_i)}{\partial B_i} = -\frac{logit(B_i) + \hat{\alpha} + \log(n_i)}{\hat{\sigma}_\alpha^2 B_i(1-B_i)} = 0 \Rightarrow \hat{B}_i = \frac{e^{-\hat{\alpha}}}{n_i + e^{-\hat{\alpha}}}, \tag{3.38}$$

$$-\ddot{\mathcal{L}}(B_i) = -\left.\frac{\partial^2 \mathcal{L}(B_i)}{\partial B_i^2}\right|_{\hat{B}_i} = \frac{1}{\hat{B}_i^2(1-\hat{B}_i)^2\hat{\sigma}_\alpha^2}. \tag{3.39}$$

We have two equations for two unknowns a_{i1} and a_{i2}. Solving for the system of equations yields the expressions for the estimated parameters $\hat{a}_{i1}$ and $\hat{a}_{i2}$:

$$\begin{aligned} \hat{\mu}_0 &= \frac{\hat{a}_{i1}}{\hat{a}_{i1} + \hat{a}_{i2}} = \hat{B}_i, \\ \hat{m} &= \hat{a}_{i1} + \hat{a}_{i2} = \frac{1}{\hat{B}_i(1-\hat{B}_i)\hat{\sigma}_\alpha^2}. \end{aligned} \tag{3.40}$$

The solutions are $\hat{a}_{i1} = \frac{1}{\hat{\sigma}_\alpha^2(1-\hat{B}_i)}$ and $\hat{a}_{i2} = \frac{1}{\hat{\sigma}_\alpha^2 \hat{B}_i}$. Now $B_i|\boldsymbol{y}$ can be approximated by $Beta(\hat{a}_{i1}, \hat{a}_{i2})$. And the c-th moment of $B_i|\boldsymbol{y}$ is

$$\hat{E}(B_i^c|\boldsymbol{y}) = \frac{B(\hat{a}_{i1} + c, \hat{a}_{i2})}{B(\hat{a}_{i1}, \hat{a}_{i2})} \tag{3.41}$$

where B stands for the beta function.

3.6.2 Posterior Distribution of Synthetic Probabilities p_{ik}^E

We approximate the unconditional posterior moments of p_{ik}^E by the conditional posterior moments with $\hat{\alpha}$ replacing α[23]:

$$E\Big((p_{ik}^E)^c|\boldsymbol{y}\Big) \approx E\Big((p_{ik}^E)^c|\hat{\alpha}, \boldsymbol{y}\Big). \tag{3.42}$$

Then, assuming the conditional posterior distribution of $(p_{i1}^E, \ldots, p_{iK}^E)$ with $\hat{\alpha}$ inserted is approximately independent Dirichlet,

$$(p_{i1}^E, \ldots, p_{iK}^E)|\hat{\alpha}, \boldsymbol{y} \sim Dirichlet(b_{i1}, \ldots, b_{iK}). \tag{3.43}$$

Then, we have

$$p_{ij}^E|\hat{\alpha}, \boldsymbol{y} = \frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}_j}}{1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}_1} + \ldots + e^{\boldsymbol{x}_i'\boldsymbol{\beta}_{K-1}}}|\hat{\alpha}, \boldsymbol{y} = \frac{G_{ij}}{G_{i1} + \ldots + G_{iK}}, \tag{3.44}$$

where $G_{ij},\ j = 1, \ldots, K$, are independent random variables following $gamma(b_{ij}, 1)$ distributions. Then,

$$e^{\boldsymbol{x}_i'\boldsymbol{\beta}_j}|\hat{\alpha}, \boldsymbol{y} = \frac{G_{ij}}{G_{iK}},\ j = 1, \ldots, K-1. \tag{3.45}$$

The mean and variance of the ratio of two independent gamma distributions are given by

$$E(e^{\boldsymbol{x}_i'\boldsymbol{\beta}_j}|\hat{\alpha}, \boldsymbol{y}) = E\left(\frac{G_{ij}}{G_{iK}}\right) = \frac{b_{ij}}{b_{iK} - 1} = \eta_{ij} \tag{3.46}$$

and

$$Var(e^{\boldsymbol{x}_i'\boldsymbol{\beta}_j}|\hat{\alpha}, \boldsymbol{y}) = Var\left(\frac{G_{ij}}{G_{iK}}\right) = \frac{\eta_{ij}(1 + \eta_{ij})}{b_{iK} - 2}. \tag{3.47}$$

From distribution (3.32), $\boldsymbol{\beta}_j$ is multivariate normally distributed with mean $\hat{\boldsymbol{\beta}}_j$ and covariance matrix $\hat{\boldsymbol{\Sigma}}_{jj}$, $j = 1, \ldots, K-1$, which is the j^{th}

$q \times q$ block in the diagonal of the covariance matrix of $\boldsymbol{\beta}$. The mean and variance of $log - normal$ distributions are easy to compute:

$$\hat{E}(e^{\boldsymbol{x}_i'\boldsymbol{\beta}_j}|\hat{\alpha}, \boldsymbol{y}) = exp(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_j + \boldsymbol{x}_i'\hat{\Sigma}_{jj}\boldsymbol{x}_i/2) = \hat{\eta}_{ij} \tag{3.48}$$

and

$$\widehat{Var}(e^{\boldsymbol{x}_i'\boldsymbol{\beta}_j}|\hat{\alpha}, \boldsymbol{y}) = \hat{\eta}_{ij}^2(e^{\boldsymbol{x}_i'\hat{\Sigma}_{jj}\boldsymbol{x}_i} - 1). \tag{3.49}$$

By matching the $(K-1)$ means in equation (3.46) and equation (3.48) and the sum of the $(K-1)$ variances in equation (3.47) and equation (3.49), we have the following system of K equations:

$$E(e^{\boldsymbol{x}_i'\boldsymbol{\beta}_j}|\hat{\alpha}, \boldsymbol{y}) = \hat{E}(e^{\boldsymbol{x}_i'\boldsymbol{\beta}_j}|\hat{\alpha}, \boldsymbol{y}),\ j = 1, \ldots, K-1 \tag{3.50}$$

and

$$\sum_{j=1}^{K-1} Var(e^{\boldsymbol{x}_i'\boldsymbol{\beta}_j}|\hat{\alpha}, \boldsymbol{y}) = \sum_{j=1}^{K-1} \widehat{Var}(e^{\boldsymbol{x}_i'\boldsymbol{\beta}_j}|\hat{\alpha}, \boldsymbol{y}). \tag{3.51}$$

Solving the above system of equations for $(b_{i1}, \ldots, b_{iK})$, we have

$$\hat{b}_{iK} = \frac{\sum_{j=1}^{K-1} \hat{\eta}_{ij}(1 + \hat{\eta}_{ij})}{\sum_{j=1}^{K-1} \hat{\eta}_{ij}^2(e^{\boldsymbol{x}_i'\hat{\Sigma}_{jj}\boldsymbol{x}_i} - 1)} + 2 \tag{3.52}$$

and

$$\hat{b}_{ij} = \hat{\eta}_{ij}(\hat{b}_{iK} - 1), j = 1, \ldots, K-1. \tag{3.53}$$

The posterior distribution of $(p_{i1}^E, \ldots, p_{iK}^E)$ is approximately a $Dirichlet(\hat{b}_{i1}, \ldots, \hat{b}_{iK})$

distribution and each $p_{ik}^E|\hat{\alpha}, \boldsymbol{y}$ is approximately a $beta(\hat{b}_{ik}, \hat{b}_{i0} - \hat{b}_{ik})$ distribution, where $\hat{b}_{i0} = \sum_{j=1}^{K} \hat{b}_{ij}$. Then,

$$\hat{E}\Big((p_{ik}^E)^c|\hat{\alpha}, \boldsymbol{y}\Big) = \frac{B(\hat{b}_{ik} + c, \hat{b}_{i0} - \hat{b}_{ik})}{B(\hat{b}_{ik}, \hat{b}_{i0} - \hat{b}_{ik})}. \tag{3.54}$$

3.6.3 Estimation of Random Effects

Under the assumption that $\boldsymbol{\beta}$ and α are independent a posteriori,

$$E(p_{ik}|\boldsymbol{y}) = \Big(1 - E(B_i|\boldsymbol{y})\Big)\bar{y}_{ik} + E(B_i|\boldsymbol{y})E(p_{ij}^E|\boldsymbol{y}), \tag{3.55}$$

$$\begin{aligned}
Var(p_{ik}|\boldsymbol{y}) &= E\Big(\frac{p_{ik}^*(1-p_{ik}^*)}{n_i + r + 1}|\boldsymbol{y}\Big) + Var\Big(p_{ik}^*|\boldsymbol{y}\Big) \\
&= E\Big(\frac{p_{ik}^*(1-p_{ik}^*)}{n_i + r + 1}|\boldsymbol{y}\Big) + Var\Big(B_i(\bar{y}_{ik} - p_{ik}^E)|\boldsymbol{y}\Big) \\
&\approx E\Big(\frac{p_{ik}^*(1-p_{ik}^*)(1-B_i)}{n_i}|\boldsymbol{y}\Big) + Var\Big(B_i(\bar{y}_{ik} - p_{ik}^E)|\boldsymbol{y}\Big) \\
&= \Big\{(1-\bar{y}_{ik})\bar{y}_{ik}[1 - E(B_i|\boldsymbol{y})] \\
&\quad + (2\bar{y}_{ik} - 1)E\Big(B_i(1-B_i)|\boldsymbol{y}\Big)\Big(\bar{y}_{ik} - E(p_{ik}^E|\boldsymbol{y})\Big) \\
&\quad + E\Big(B_i^2(1-B_i)|\boldsymbol{y}\Big)E\Big((\bar{y}_{ik} - p_{ik}^E)^2|\boldsymbol{y}\Big)\Big\}/n_i \\
&\quad + Var\Big(B_i(\bar{y}_{ik} - p_{ik}^E)|\boldsymbol{y}\Big).
\end{aligned} \tag{3.56}$$

The approximation in equation (3.56) is a first-order Taylor approximation. By inserting the moment estimates of the shrinkage B_i and the moment estimates of the expected probability p_{ik}^E as presented in equation (3.41) and the equation (3.54) into the equation (3.55) and the

equation (3.56), we obtain the estimated posterior mean and posterior variance of the multinomial probabilities, respectively.

3.7 Conclusion

This chapter develops the ADM to estimate the multinomial probabilities in a multinomial-Dirichlet-logit model. The procedure starts from adjusting the likelihood of the hyper-parameters by a third level hyper-prior on the hyper-parameters. This adjustment removes the possibility that the variance component r estimate occurs at an infinite value. The REML adjustment to the posterior distribution of the hyper-parameters corrects the bias in the variance component estimate. The adjustment to the determinant of the Hessian matrix speeds up the computation. When estimating the multinomial probabilities, we introduce the variance in the hyper-parameter estimates. In Chapter 4, we discuss results from some simulation studies to check the performance of the proposed estimates.

Chapter 4: Comparisons with Other Methods

4.1 Introduction

This chapter demonstrates the advantages of the ADM proposed in Chapter 3 over other parameter estimation procedures, including sampling-based approaches and empirical Bayes methods. As the previous studies in ADM do, our proposed approach also provides some attractive features. First, the proposed ADM is designed to provide reasonable point estimates and interval estimates for all the parameters for all N observations, unlike some other methods. Second, although MCMC can also provide a full range of inferences, the computation speed of our method is hundreds of times faster than the MCMC procedure through RStan[52] and thus permits its evaluation by repeated use in simulations. Third, simulation studies have shown that the ADM has better operating characteristics than the EB methods.

Section 4.2 compares estimates from our method with the corresponding estimates from the MCMC approach. The comparison with MCMC ensures the accuracy of the ADM estimates and emphasizes the overwhelmingly fast speed of the ADM. In Section 4.3, we compare the inferences from the ADM with two alternative EB methods. The comparison with the two EB methods

verifies the ill behavior of the MLE for the variance component and empirically demonstrates the better operating characteristics of the ADM estimates. Finally, Section 4.4 presents a summary and discussion.

4.2 Comparison with MCMC

To make a newly proposed procedure useful, it must be accurate in estimation. Thus, comparing the ADM estimates with the MCMC estimates will be of interest. The data used in this study is a small random sample of 10 groups from a race count data extracted from Twitter big data. Each observation in the sample contains non-integer counts for five race categories in a small area and the information of its state. The details about the Twitter data are given in Chapter 5. This sample is used only for illustration and to verify the accuracy of the estimates obtained from the ADM for a small sample. The analysis of the complete Twitter data is given in Chapter 5. Since MCMC implemented by RStan can only deal with integer counts, we first round the counts in the sample to the nearest integers $(y_{i1}, \ldots, y_{i5})$ in Table 4.1 to feed into the MCMC package Rstan[52]. The variable x_{i1} is the code for state (1: California; 0: Florida; -1: Texas). To be consistent, we also run the ADM on the rounded sample data although the ADM can handle non-integer counts. The results from the two approaches are presented in Table 4.1. The estimates of both the hyperparameters $\boldsymbol{\beta}$ and r and the multinomial probabilities $\boldsymbol{p}_i$ are close to the corresponding estimates from MCMC. The standard errors for

the multinomial probability estimates from the two procedures are also close. But the difference between the speeds of the two methods on this sample data is obvious. On the same laptop, the ADM is hundreds of times faster than the MCMC approach (burns-in of 5000 samples, 5000 samples after burns-in and 4 chains starting from 4 different sets of initial hyper-parameter values) using the Rstan package[52]. The ADM directly gives the formulas for the point estimates and the interval estimates of all the parameters, avoiding the trouble of convergence checking as in the MCMC approach. One more advantage of the ADM is that each time a model is applied to the same dataset, the estimated results will be the same; while MCMC spits out different estimates each time it runs on the same dataset if seed is not set. This randomness can be awkward for some legal and public policy applications[13].

Table 4.1: Comparison of Estimates Generated by MCMC and ADM

							MCMC					ADM				
	Data						$\hat{\boldsymbol{\beta}} = \begin{pmatrix} 2.620 & -0.207 \\ 1.102 & -0.190 \\ 0.524 & -0.022 \\ 1.535 & 0.071 \end{pmatrix}, \hat{r} = 43.975$					$\hat{\boldsymbol{\beta}} = \begin{pmatrix} 2.541 & -0.201 \\ 1.051 & -0.177 \\ 0.499 & -0.019 \\ 1.470 & 0.071 \end{pmatrix}, \hat{r} = 42.660$				
obs i	y_{i1}	y_{i2}	y_{i3}	y_{i4}	y_{i5}	x_{i1}	$\hat{p}_{i1}$	$\hat{p}_{i2}$	$\hat{p}_{i3}$	$\hat{p}_{i4}$	$\hat{p}_{i5}$	$\hat{p}_{i1}$	$\hat{p}_{i2}$	$\hat{p}_{i3}$	$\hat{p}_{i4}$	$\hat{p}_{i5}$
1	289	62	45	108	19	1	0.551 (0.022)	0.119 (0.014)	0.086 (0.012)	0.208 (0.018)	0.037 (0.008)	0.549 (0.021)	0.119 (0.014)	0.086 (0.012)	0.208 (0.017)	0.037 (0.008)
2	261	65	46	187	19	1	0.457 (0.020)	0.113 (0.013)	0.080 (0.010)	0.317 (0.019)	0.034 (0.007)	0.456 (0.020)	0.113 (0.013)	0.080 (0.011)	0.317 (0.019)	0.034 (0.007)
3	2	0	1	4	1	1	0.474 (0.087)	0.098 (0.051)	0.086 (0.047)	0.281 (0.077)	0.062 (0.038)	0.470 (0.082)	0.102 (0.047)	0.090 (0.041)	0.275 (0.077)	0.063 (0.037)
4	233	45	19	58	13	0	0.625 (0.024)	0.123 (0.016)	0.054 (0.011)	0.162 (0.018)	0.036 (0.009)	0.626 (0.024)	0.123 (0.016)	0.054 (0.011)	0.161 (0.018)	0.036 (0.009)
5	172	41	10	43	9	0	0.619 (0.028)	0.146 (0.020)	0.041 (0.011)	0.161 (0.021)	0.034 (0.010)	0.617 (0.027)	0.146 (0.020)	0.041 (0.011)	0.161 (0.021)	0.034 (0.010)
6	159	28	12	19	8	0	0.682 (0.029)	0.124 (0.021)	0.056 (0.014)	0.102 (0.019)	0.036 (0.012)	0.681 (0.029)	0.124 (0.020)	0.056 (0.014)	0.101 (0.018)	0.037 (0.012)
7	7050	1589	712	1966	373	-1	0.603 (0.005)	0.136 (0.003)	0.061 (0.002)	0.168 (0.003)	0.032 (0.002)	0.603 (0.005)	0.136 (0.003)	0.061 (0.002)	0.168 (0.003)	0.032 (0.002)
8	862	155	65	115	39	-1	0.694 (0.013)	0.126 (0.010)	0.053 (0.006)	0.095 (0.008)	0.032 (0.005)	0.694 (0.013)	0.126 (0.009)	0.053 (0.006)	0.095 (0.008)	0.032 (0.005)
9	149	31	17	155	9	-1	0.434 (0.027)	0.091 (0.015)	0.049 (0.011)	0.400 (0.026)	0.026 (0.008)	0.432 (0.025)	0.091 (0.014)	0.049 (0.011)	0.401 (0.024)	0.027 (0.008)
10	3	1	2	0	0	-1	0.592 (0.082)	0.137 (0.055)	0.099 (0.049)	0.138 (0.057)	0.034 (0.028)	0.587 (0.073)	0.138 (0.050)	0.100 (0.064)	0.139 (0.057)	0.036 (0.027)

4.3 Comparison with Empirical Bayes Methods

We also conduct a Monte Carlo simulation study comparing inferences from the ADM with those from two alternative empirical Bayes (EB) methods: EB-MLE and EB-REML. Five datasets with different combinations of (N, K, q) are used in the simulation study. All scenarios are based on the Twitter race count data. The complete Twitter dataset has a wide range of group sizes (n_i), from 1 to 51170. The covariate matrix $\boldsymbol{X}$ (with two covariates: the intercept and the state) and the group sizes n_i, $i = 1, \ldots, N$, for the five experiments are all sampled from the complete Twitter data. The simulated data and the two alternative methods, EB-MLE and EB-REML, are introduced in Section 4.3.1 and Section 4.3.2, respectively. Section 4.3.3 summarizes the operating characteristics of the ADM and the two EB procedures.

4.3.1 Simulated Data

There are five datasets with various combinations of (N, K, q) used in this simulation study, all sampled from the complete Twitter dataset. Only the covariate matrix $\boldsymbol{X}$ and the group sizes n_i, $i = 1, \ldots, N$, in the Twitter sample are used. Different known values for the hyper-parameters $\boldsymbol{\beta}$ and r are assigned to the five datasets. Please refer to Method d in Table 4.2 (Column: avg. $\hat{r}$ and Column: avg. $\hat{\boldsymbol{\beta}}$) for the true values of the hyper-parameters for the five datasets. For

each dataset, we first generated 100 replicates of probability vectors $\boldsymbol{p}_i \sim Dirichlet(rp_{i1}^E, \ldots, rp_{iK}^E)$, $i = 1, \ldots, N$, and then generated 100 replicates of counts $\boldsymbol{y}_i \sim Multinomial(n_i, \boldsymbol{p}_i)$, $i = 1, \ldots, N$. We ran the three approaches on each of the 100 simulated datasets generated with the same $\boldsymbol{X}$, n_i and hyper-parameters. We computed the averages of the $100 \times N$ shrinkages and the 100 sets of hyperparameter estimates for 100 simulated datasets. The results are displayed in the last three columns of Table 4.2.

4.3.2 Alternative Methods

Empirical Bayes methods are based on the first two levels as in (3.1) and (3.2) of our three level model. The means and the variances of the multinomial parameters are estimated using equation (3.10) and equation (3.11) by plugging in the ML or REML estimates for the hyper-parameters. In EB-MLE, the hyper-parameters are estimated by maximum likelihood method and in EB-REML, the hyper-parameters are estimated by maximum likelihood method on the REML corrected likelihood. The two sets of estimates for the hyper-parameters are referred to as the MLE estimates and the REML estimates.

4.3.3 Operating Characteristics

This subsection reports results on different operating characteristics from our Monte Carlo simulation study. For illustration, we only report results for multinomial probabilities for the first category $p_{i1},\ i = 1, \cdots, N$.

Table 4.2 displays results for average operating characteristics, where the average is over all the N groups. The second column in Table 4.2 displays the average coverage rates for the three estimation methods. The coverage probability for group i is defined as

$$\Pr(\hat{p}_{i1,0.025} < p_{i1} < \hat{p}_{i1,0.975} | \beta, r), \tag{4.1}$$

where Pr denotes the probability with respect to the joint distribution of $(y_i,\ p_i,\ i = 1, \cdots, N)$. We approximate this probability using the 100 replicates generated from the multinomial-Dirichlet-logit model given fixed values of the hyper-parameters β and r. For the two EB methods, the endpoints $\hat{p}_{i1,0.025}$ and $\hat{p}_{i1,0.975}$ are the 2.5% and 97.5% quantiles of the distribution:

$$Beta(y_{i1} + \hat{r}\hat{p}^E_{i1}, (n_i - y_{i1}) + \hat{r}(1 - \hat{p}^E_{i1}), \tag{4.2}$$

respectively. In equation (4.2), $\hat{r}$ is the MLE estimate and the REML

estimate of r for the EB-MLE method and the EB-REML method, respectively; and $\hat{p}_{i1}^{E}$ is computed by inserting the ML and REML estimates of $\boldsymbol{\beta}$ for the EB-MLE method and the EB-REML method, respectively. For the ADM, the endpoints $\hat{p}_{i1,0.025}$ and $\hat{p}_{i1,0.975}$ are the 2.5% and 97.5% quantiles of the Beta distribution:

$$Beta(\hat{b}_{i1}, \hat{b}_{i2}), \tag{4.3}$$

where $\hat{b}_{i1}$ and $\hat{b}_{i2}$ are computed from the estimates $\hat{p}_{i1} = \hat{E}(p_{i1}|\boldsymbol{y})$ and $\hat{\sigma}_{i1}^2 = \widehat{Var}(p_{i1}|\boldsymbol{y})$ by solving the following system of equations

$$\begin{aligned} \hat{p}_{i1} &= \frac{\hat{b}_{i1}}{\hat{b}_{i1} + \hat{b}_{i2}} \ , \\ \hat{\sigma}_{i1}^2 &= \frac{\hat{b}_{i1}\hat{b}_{i2}}{(\hat{b}_{i1} + \hat{b}_{i2})^2(\hat{b}_{i1} + \hat{b}_{i2} + 1)}. \end{aligned} \tag{4.4}$$

When computing the endpoints for the ADM, the uncertainty in the estimates of the hyper-parameters r and $\boldsymbol{\beta}$ is taken into consideration. This method of computing the coverage rate is named Rao-Blackwellization and has improved the simulation accuracy substantially[13] under the condition when only 100 trials are available for each i. Compute the average of all $100 \times N$ coverage rates as the average coverage rate for a procedure. From Table 4.2, ADM has better average simulated coverage rate than the EB procedures. The under-coverage rates for the EB methods can be explained by the fact that (1) occa-

sionally r is estimated at ∞ making the variance (3.11) estimated at 0 and (2) the naive EB methods do not take into consideration the uncertainty in the hyper-parameter estimates when assessing the posterior variances. Even a method achieves the nominal average coverage rate 0.95, the coverage rate for the method may vary over the groups depending on the group sizes or other predictor variables. This is an undesirable feature of a method. Figure 4.1 plots simulated coverage rates against groups arranged in order of group sizes. We observe that the coverage rate increases with group sample size (n_i) for the two EB methods. In contrast, the simulated coverage rates based on the ADM are much more stable across different group sizes and stay around the ideal 0.95 coverage rate. Our proposed ADM provides intervals that are wider than those generated by the two EB methods (see Table 4.2). These wider intervals, caused by the inclusion of additional uncertainty due to the hyper-parameters by the ADM, are partly responsible for its better coverage. For the case $(N = 30, K = 5, q = 2)$ in Table 4.2, the proposed ADM provides better coverage than EB-MLE method even though its average interval width is narrower than that of the EB-MLE method.

The loss function used to compare estimators is the sum of the squared error losses,

$$L(\hat{\boldsymbol{p}}, \boldsymbol{p}) = \sum_{i=1}^{N} \|\hat{\boldsymbol{p}}_i - \boldsymbol{p}_i\|^2 , \tag{4.5}$$

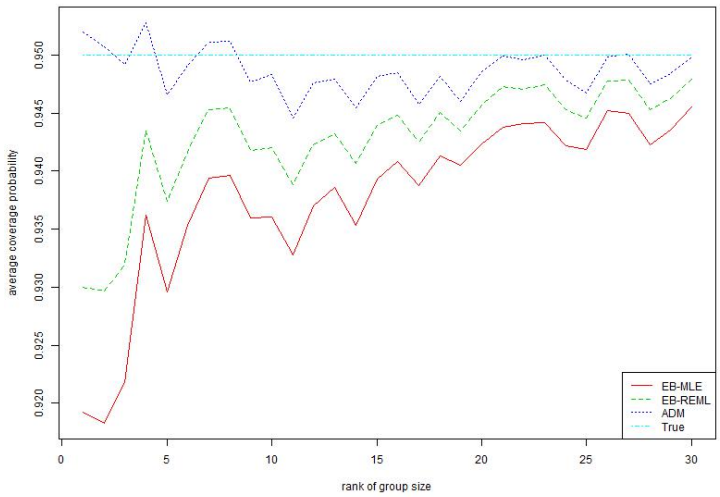


Figure 4.1: Average coverage rate of 100 replicates vs. group index, N=30, K=5, q=2. Group size increases from 10 to 139 from group 1 to group 30. The average coverage rates for EB-MLE, EB-REML and ADM for the complete data are 0.938,0.943 and 0.949, respectively.

where $\hat{\boldsymbol{p}} = \{\hat{\boldsymbol{p}}_i\}$ and $\boldsymbol{p} = \{\boldsymbol{p}_i\}$, $i = 1, \ldots, N$, are the sets of estimated multinomial probabilities and true multinomial probabilities for the 100 replicates, respectively. We compare risk or total mean squared error (MSE), over all the groups, defined as $E[L(\hat{\boldsymbol{p}}, \boldsymbol{p})]$, where the expectation is taken over the joint distribution of $(\boldsymbol{y}_i, \boldsymbol{p}_i,\ \ i = 1, \cdots, N)$ in the multinomial-Dirichlet-logit model for given hyper-paraters β and r. This is also a reasonable evaluation criterion under the classical EB approach. The optimum estimator for this loss function is the mean $E(\boldsymbol{p}_i|data)$. Thus, p^*_{ik} given by (3.10) is the best estimator when $\boldsymbol{\beta}$ and r are known - method labeled (d) in Table 4.2 and gives a lower bound to what is realistically possible. The risks, estimated as the average of these 100 losses, are given in the column with heading Risk of Table

4.2. The ADM estimator of $\{\boldsymbol{p}_i\}$ has equally good or better risks than the EB estimators and is stable in performance for the five datasets, especially for the dataset with $(N = 30, K = 5, q = 2)$.

The average of estimates of r over 100 replicates (Column: $avg.\ \hat{r}$) by the EB-MLE and the EB-REML in the first and the fifth examples can occasionally occur at an infinite value, but the ADM does not encounter such a problem for all five cases. We also observe that REML adjustment effectively corrects the bias in the estimation of r. The average of the estimates of $\boldsymbol{\beta}$ (Column: $avg.\ \hat{\boldsymbol{\beta}}$) are relatively insensitive to the average of r, because r and $\boldsymbol{\beta}$ are relatively independent.

4.4 Discussion

This chapter demonstrates the advantages of the proposed ADM in parameter estimation over the other commonly used approaches, including the MCMC approach and two empirical Bayes methods. In our simulation experiment, the ADM shows multiple advantages over the MCMC. Firstly, the computation speed of the proposed ADM is hundreds of times faster than MCMC without sacrificing the accuracy of the estimates. Secondly, the estimates stay the same each time the ADM for a model is applied to the same dataset as observed in previous studies[13;38]. This characteristic of the ADM is favored by some data practitioners and cannot be achieved by the sampling-based procedures. Through the comparisons with the empirical Bayes methods, our method al-

Table 4.2: Operating Characteristics: Coverage Rate, Interval Width and Risk

Method	Coverage	Interval	*avg.* $\hat{\sigma}_1$	Risk	$\hat{B}$	*avg.* $\hat{r}$	*avg.* $\hat{\boldsymbol{\beta}}$
N=3, K=5, q=2							
a EB-MLE	0.710	0.062	0.016	0.009	0.395	150.34*	$\begin{pmatrix} 3.16 & -0.26 \\ 1.47 & -0.93 \\ 0.88 & 0.36 \\ 1.91 & 0.79 \end{pmatrix}$
b EB-REML	0.905	0.083	0.021	0.006	0.156	42.07*	$\begin{pmatrix} 2.91 & -0.23 \\ 1.27 & -0.87 \\ 0.75 & 0.31 \\ 1.71 & 0.76 \end{pmatrix}$
c ADM	0.941	0.090	0.023	0.006	0.085	36.66	$\begin{pmatrix} 2.69 & -0.20 \\ 1.12 & -0.80 \\ 0.66 & 0.29 \\ 1.55 & 0.73 \end{pmatrix}$
d Ideal	0.95	0.090	0.023	0.005	0.085	30	$\begin{pmatrix} 2.80 & -0.30 \\ 1.20 & -0.90 \\ 0.70 & 0.20 \\ 1.60 & 0.70 \end{pmatrix}$
N=6, K=5, q=2							
a EB-MLE	0.927	0.098	0.025	0.045	0.145	20.13	$\begin{pmatrix} 0.55 & -0.71 \\ 1.33 & -0.84 \\ 0.07 & 0.38 \\ -0.47 & 1.04 \end{pmatrix}$
b EB-REML	0.939	0.101	0.026	0.043	0.114	13.65	$\begin{pmatrix} 0.51 & -0.66 \\ 1.26 & -0.79 \\ 0.07 & 0.35 \\ -0.44 & 0.98 \end{pmatrix}$
c ADM	0.944	0.104	0.027	0.042	0.102	11.78	$\begin{pmatrix} 0.49 & -0.64 \\ 1.23 & -0.76 \\ 0.07 & 0.34 \\ -0.42 & 0.94 \end{pmatrix}$
d Ideal	0.95	0.105	0.027	0.040	0.093	10	$\begin{pmatrix} 0.50 & -0.70 \\ 1.20 & -0.80 \\ 0.10 & 0.30 \\ -0.40 & 0.90 \end{pmatrix}$
N=15, K=3, q=2							
a EB-MLE	0.939	0.117	0.030	0.059	0.160	26.74	$\begin{pmatrix} 0.54 & 0.27 \\ -0.45 & 2.49 \end{pmatrix}$
b EB-REML	0.943	0.120	0.031	0.058	0.149	23.06	$\begin{pmatrix} 0.54 & 0.27 \\ -0.44 & 2.45 \end{pmatrix}$
c ADM	0.949	0.126	0.032	0.058	0.141	20.65	$\begin{pmatrix} 0.53 & 0.26 \\ -0.43 & 2.42 \end{pmatrix}$
d Ideal	0.95	0.121	0.031	0.051	0.141	20	$\begin{pmatrix} 0.50 & 0.24 \\ -0.46 & 2.47 \end{pmatrix}$
N=15, K=3, q=3							
a EB-MLE	0.946	0.070	0.018	0.035	0.055	13.07	$\begin{pmatrix} 1.19 & 0.03 & -0.17 \\ 0.31 & 2.11 & 0.96 \end{pmatrix}$
b EB-REML	0.947	0.070	0.018	0.035	0.048	11.09	$\begin{pmatrix} 1.16 & 0.02 & -0.16 \\ 0.30 & 2.05 & 0.93 \end{pmatrix}$
c ADM	0.949	0.071	0.018	0.035	0.045	10.18	$\begin{pmatrix} 1.15 & 0.01 & -0.16 \\ 0.30 & 2.02 & 0.91 \end{pmatrix}$
d Ideal	0.95	0.071	0.018	0.034	0.044	10	$\begin{pmatrix} 1.20 & 0.00 & -0.19 \\ 0.30 & 2.00 & 0.91 \end{pmatrix}$
N=30, K=5, q=2							
a EB-MLE	0.937	0.108	0.017	0.274	0.082	5.59*	$\begin{pmatrix} 0.72 & 0.77 \\ 1.22 & 1.85 \\ 2.11 & 0.98 \\ -1.15 & 1.07 \end{pmatrix}$
b EB-REML	0.940	0.069	0.018	0.133	0.042	5.29*	$\begin{pmatrix} 0.73 & 0.77 \\ 1.23 & 1.86 \\ 2.13 & 0.97 \\ -1.15 & 1.07 \end{pmatrix}$
c ADM	0.950	0.071	0.018	0.102	0.032	5.14	$\begin{pmatrix} 0.72 & 0.77 \\ 1.22 & 1.86 \\ 2.12 & 0.97 \\ -1.15 & 1.07 \end{pmatrix}$
d Ideal	0.95	0.070	0.018	0.101	0.031	5	$\begin{pmatrix} 0.69 & 0.74 \\ 1.19 & 1.83 \\ 2.08 & 0.94 \\ -1.12 & 1.07 \end{pmatrix}$

*The average is infinite. This value is the median.

ways generates a finite estimate for the hyper-parameter r even for a small N and the multinomial probability estimates obtained through the ADM have better operating characteristics.

The proposed ADM includes a series of adjustments to the posterior distribution of the hyper-parameters and to some other posterior distributions. The steps for the ADM for a multinomial-Dirichlet-logit model has been discussed in Chapter 3. The successful performance of the ADM, as discussed above in this section, relies on a series of adjustments to the posterior distribution of the hyper-parameters and the ADM approximations to the posterior distributions of the shrinkage factors and the posterior distributions of the synthetic proportions. Firstly, the exact posterior distribution of the hyper-parameters is carefully adjusted so that adjusted posterior modes of the hyper-parameters are always in the interior of the parameter space. This very first adjustment to the posterior distribution of the hyper-parameters corrects the ill-behavior in the MLE or the REML of the variance component estimate. Secondly, the REML type correction to the posterior distribution of the hyper-parameters reduces the bias in the variance component estimate, although the REML type correction alone does not eliminate occasional infinite value for the variance component estimate. Thirdly, the adjustment of approximating the determinant of the Hessian matrix by its diagonal terms in the REML adjusted posterior distribution of the hyper-parameters simplifies and speeds up the computation. The first two adjustments to the posterior distribution of the

hyper-parameters improve the approximation to the distribution of the hyper-parameters for a small sample. The ADM approximations to the posterior distributions of the shrinkage factors and the synthetic probabilities make the approximations to the posterior means and posterior variances of the multinomial probabilities in closed-forms possible, avoiding sampling. The approximations in closed-forms save computational cost significantly. The computation of the posterior means and posterior variances of the multinomial probabilities incorporates the variance in the hyper-parameter estimates through the estimates of the moments of the shrinkage factors and the synthetic probabilities in the proposed procedure. The resulting interval estimates of the multinomial probabilities in our proposed procedure are generally wider than those from the EB-plugin procedures. The wider interval estimates partly explain the higher coverage rates in the simulation studies.

Chapter 5: Application

5.1 Introduction

In the previous chapters, we have either reviewed or developed the ADM to estimate the parameters in a series of hierarchical Bayes models, including the Poisson-gamma, the normal-normal, the binomial-beta and the multinomial-Dirichlet regression models. This chapter will give two data analysis examples using the hierarchical Bayes models. In the first example, we propose an alternative method to the ACS direct estimation method in calculating the point estimates and the interval estimates for the small area gender proportions. We introduce the Twitter direct gender proportion estimate as a covariate in the binomial-beta logit model to analyze the American Community Survey (ACS) small area gender count data. And the proposed method reduces the margins of error for the small area proportion estimates. The second example analyzes the Twitter non-integer small area race counts using the multinomial-Dirichlet-logit model. The small area race proportion estimates generated by this model are weighted sums of the direct estimates and the synthetic estimates from the regression. The binomial-beta-logit model and the multinomial-Dirichlet-logit model are implemented using the ADM as introduced in Chapter 2 and Chap-

ter 3, respectively.

5.2 Small Area Gender Distribution

5.2.1 Introduction

The ACS is an annual nationwide survey in the United States. The ACS data is used to allocate more than $675 billion in federal and state funds[6]. Different stakeholders (e.g., public officials, planners and entrepreneurs) rely on ACS data for timely precise estimates of different socio-economic characteristics of its people at the national and different small area levels.

Every year the Census Bureau releases ACS one-year and five-year Public Use Microdata Samples (PUMS) on its website. The PUMS contain records about individual people or housing units from all the states in the nation[7]. The five-year PUMS data is obtained by merging five one-year PUMS databases. National estimates can be computed from these two sets of data, although there is a trade-off between the quality and the timeliness of the estimates when using these two sets of data. The five-year PUMS data contains enough sample points to generate reliable five-year period estimates. These estimates are, however, not appropriate for the most recent time frame. On the other hand, there are fewer sample points in the one-year PUMS data compared with the

five-year PUMS data. This results in estimates with large margins of error using the method given by the Census Bureau[4;54].

As an illustrative example, the margins of error for the direct survey-weighted estimates of proportions of males ($r_i,\ i = 1, \ldots, 2378$) in the year 2016 in all the 2,378 Public Use Microdata Areas (PUMAs)[3] are computed. The PUMAs are statistical geographic areas defined by the Census Bureau in the year 2010. The histogram of the margins of error is presented in Figure 5.1(a). The margins of error are computed using the method documented by the Census Bureau[4]. This method is based on the final weights and the 80 weight replicates for each record. In the 2016 one-year PUMS data, male and female are coded as 1 and 2, respectively. For computational convenience and notational simplicity, we recode the female as 0. Then, the Census Bureau estimates of the number of males y_i and the proportion of males r_i in the i^{th} PUMA are respectively given by

$$y_i = \sum_{j=1}^{n_i} w_{ij} y_{ij} \tag{5.1}$$

and

$$r_i = y_i / \sum_{j=1}^{n_i} w_{ij}, \tag{5.2}$$

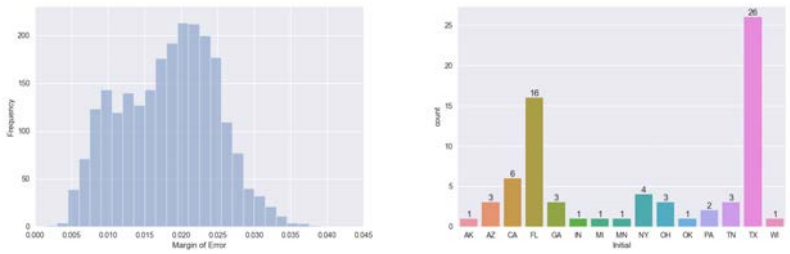
where n_i is the sample size (e.g., the number of individual records in the sample) in the i^{th} PUMA, y_{ij} is the indicator value of being male or not and w_{ij} is the weight for the j^{th} individual in the i^{th} PUMA.

The estimated margin of error for the survey-weighted estimate of the male proportion in each PUMA is given by (for notational simplicity, the subscript i is dropped in this equation since the equation is applied to all the PUMAs)

$$ME(r) = 1.96\sqrt{\frac{4}{80}\sum_{k=1}^{80}(r_k - r)^2}, \tag{5.3}$$

where r is the male proportion computed using the final weights and the 80 r_k's are the direct survey-weighted estimates of the male proportions computed using the 80 weight replicates. From Figure 5.1(b), Florida and Texas are the two states that have the most PUMAs with margins of error greater than the cutoff 3%. To solve this problem, in this example we apply the ADM using a binomial-beta-logit model that combines information from data derived from Twitter and ACS PUMS data. We treat the weighted non-integer counts of males in the i^{th} PUMA as the observed count y_i, the number of individual records in the i^{th} PUMA as n_i and the Twitter male proportion in the i^{th} PUMA as the covariate x_{i1} in the binomial-beta-logit model (2.82). The goal is to obtain estimate of p_i, the proportion of males, and the associated variance estimate.

With the popularity of social media[50] and the tracking technology Global Positioning System (GPS) embedded in mobile devices[49] all around the world, a large amount of data about social media users



(a) *(b)*

Figure 5.1: Estimated margins of error for male proportion estimates (year: 2016). (a) Histogram of estimates of margins of error of direct estimates in the 2,378 PUMAs; (b) Number of PUMAs with margins of error greater than 3% in selected states. The continental states not in this figure have 0 PUMA with margin of error greater than 3%.

is accumulating. The information about social media users in this big data set can include the location and the profile (e.g., gender, age, occupation and interest) of each individual. Although at the current stage, there are still a lot of open questions about this big dataset, more and more researchers are joining in and conducting studies to get an ever-improving picture of the social media population. There is already some work related to the social media population: (1) study of the time evolution of social media demographics and comparisons with the results from the U.S. Census[35]; (2) improvement of the accuracy of location estimates of social media users[11;12;26;27;31;43]; (3) strengthening the capability of inferring social media users' age, occupation and socio-economic status[40;41;46]; and (4) application of statistical methods to correct bias in the social media population[33]. It can be predicted that with more sophisticated techniques developed in understanding the so-

cial media users' profiles, it is possible to get enriched and reliable sets of distributions of different aspects about the social media population. Thus, it is meaningful to develop useful statistical models to link the social media estimates and the traditional data (e.g., ACS data). This type of research is promising to reduce survey cost and to provide better estimates. Thus, a growth in model-based small area estimation has been observed and has been applied to different applications of social sciences, including (1) poverty estimation[15;36;39], (2) labor force estimation[10;30] and most recently (3) literacy rate estimation[44]. Schmid et al.[44] adopt the mobile phone data combined with survey data to estimate literacy rate when census data is missing in developing countries. This chapter develops model-based small area gender proportion estimates and their associated margins of error at the PUMA level by using the ADM introduced in Chapter 2 implemented on a binomial-beta-logit model. This methodology can be extended to estimation at any geographic granularity (e.g. state, county, city) whenever there are enough quality data points or reliable derived variables from alternative data resources, which can serve as the auxiliary variables in the statistical model. This work differs from other papers that use big data in small area estimation. First, we use a discrete binomial model on non-integer survey-weighted counts, unlike normality-based models assumed by other researchers. Secondly, we use the ADM as introduced in Chapter 2 instead of classical EB methods used by others.

This section is organized as follows. Section 5.2.2 introduces the data collection method and the variables contained in the two datasets: Twitter data and 2016 ACS one-year PUMS data, used in this work. Section 5.2.3 describes the data processing and information extraction procedures and the binomial-beta-logit model. Section 5.2.4 presents the improved estimates with smaller margins of error from the model. Section 5.2.5 summarizes the contributions of this work and discusses possible application scenarios and possible future research in theoretical model development. The development of ADM for more distributions can make the proposed method fit into a wider set of data with a range of distributions.

5.2.2 Data

There are two major datasets used in this work:

Twitter Data: This dataset is collected using Twitter Streaming API, which returns 1% of real-time Tweets with the location filter set to be bounded by the latitude-longitude box [124.7625, 66.9326]W $\times$ [24.5210, 49.3845]N[33]. This dataset contains 161,771,878 Twitter messages sent by 3,670,604 active Twitter users between July 10, 2017 and October 20, 2017 in the continental United States (excluding Alaska, Hawaii, and offshore US territories and possessions). The Twitter users in the sample amount to more than 1.13 percent of the US population, which

was 323,127,513 by July 1, 2016[5]. Each tweet in the sample is composed of information about the message (e.g., the content of the message, the time the tweet is posted and the geo-tag) and the user (e.g., the self-reported profile including name, location and company). And we note that the Twitter users do not represent the US population at the time we collect the data. Children and the elderly are unlikely to use Twitter, and the teens and the young adults are more likely to use it. It is also possible that one user has multiple accounts and some profiles may be partly falsified.

2016 ACS Data: The ACS 2016 PUMS data is downloaded from the U.S. Census Bureau website[8]. PUMS data provides anonymized individual responses to questionnaire with variables covering different aspects (e.g., social status, economic status, housing and demographics). Each record in the data with either household or individual as the unit has a final weight and 80 replicate weights. The final weights are used to compute the estimates as in equations (5.1) and (5.2) and the replicate weights are used to compute estimates of margins of error as in equation (5.3).

5.2.3 Methods

To obtain the gender distribution in each PUMA, we must first assign gender and PUMA to each Twitter user based on the information in

the tweets; then we can count the male in each PUMA.

Gender: We first extract the self-reported name from each tweet and use the first name to infer about the gender of the Twitter user. To get the name lists that represent male and female, respectively, we aggregate the 1,000 most popular baby names in each year from 1918 to 2017 (100 years) from the Social Security website[45]. There are 2,774 male and 3,546 female names in the lists. As 503 names occur in both lists, we remove 273 names with no separating power (e.g., shows up equally often as male and female names), resulting in 2,397 male and 3,147 female names, which can be used to estimate a Twitter user's gender. We observe that there is a match for 43.7% of the users in the sample and there are 836,510 identifiable males and 768,427 identifiable females. That is, 52.1% male and 47.9% female in the sample. This proportion matches the nationwide Twitter gender distribution[48].

Location: Location information can be inferred from the geotags contained in each tweet. A twitter user in our data can post multiple times from different locations, which results in a set of geotags for a single user. To pair a user with only one location, we first aggregate all the geotags for a single Twitter user and then associate the user's most frequent geotag with the user. Based on our data, 3,521,887 Twitter users (about 96% of all sample users) in the sample are geo-tagged. From Figure 5.1(b), Florida and Texas are the two states that have the

most PUMAs with margins of error greater than the cutoff 3%. For illustrative purposes and low computational cost, we only selected users from these two states based on the state information in the geotag. We fed the geotag to Bing Maps API[34] to get the longitude and latitude for each user. There were 539,388 sample tweets in total in these two states that were decoded successfully by API. Then we use the PUMA shapefile[53] (e.g., by checking whether the point longitude and latitude of a user are inside the boundary of a PUMA) to assign a PUMA to each Twitter user. There are 44,557 sample users who are successfully geo-identified, in the PUMAs of large margins (e.g., greater than 3%) in the two states.

Next we selected the 30 PUMAs with highest margins of error in Florida and Texas and the Twitter users identified as being located within these PUMAs. Finally, we summarized the Twitter male proportion (Column: Twitter), the sample size (Column: Size) in the PUMS data, the weighted male count ($\hat{y}_i$, Column:Male) and the margin of error for direct male proportion estimate (Column:Margin) in each of the 30 PUMAs in Table 5.1. The weighted male count $\hat{y}_i$ is calculated using the normalized weights following equation (5.4)

$$\hat{y}_i = \sum_{j \in s_i} w^*_{ij} y_{ij} \ , \tag{5.4}$$

where

$$w_{ij}^{*} = \frac{w_{ij}}{\sum_{j \in s_i} w_{ij}} n_i \tag{5.5}$$

is the normalized weight, s_i stands for the sample in the i^{th} PUMA, w_{ij} is the weight for the j^{th} sample in the i^{th} PUMA, n_i is the sample size in the i^{th} PUMA, and y_{ij} is the indicator variable for being male. The table is in descending order of the column Margin, which is the margin of error for each PUMA.

Model: We model the weighted male counts ($\hat{y}_i$, Column: Male in Table 5.1) in each of the 30 PUMAs by a binomial-beta-logit model. The model has three levels as described in Section 2.4, Chapter 2:

$$\hat{y}_i | p_i \sim Binomial(n_i, p_i), \tag{5.6}$$

$$p_i | \boldsymbol{\beta}, r \sim Beta(r p_i^E, r(1 - p_j^E)) \ , \tag{5.7}$$

$$\boldsymbol{\beta} \sim Uniform \ on \ \boldsymbol{R}^2, \ \ 1/r \sim Uniform(0, \infty), \tag{5.8}$$

where $\hat{y}_i$ is the weighted number of males out of n_i records in the i^{th} PUMA and p_i^E is the synthetic estimate of the random effect p_i, defined as:

$$p_i^E = E(p_i | \boldsymbol{\beta}, r) = \frac{e^{\boldsymbol{x_i}^T \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x_i}^T \boldsymbol{\beta}}}, \tag{5.9}$$

for $i = 1, \cdots, N$. The vector $\boldsymbol{x_i}$ contains the intercept term ($x_{i0} = 1$) and the covariate - the Twitter male proportion (x_{i1}). The logis-

Table 5.1: Twitter Data for the 30 PUMAs with Largest Margins of Error

State	PUMA	Male*	Size*	**Margin*** (%)	**Twitter*** (%)
Florida	9501	562.61	1164	3.39	62.43
Florida	8604	497.51	972	3.38	55.57
Florida	1114	527.33	1093	3.38	55.22
Texas	2312	624.13	1172	3.36	48.36
Texas	2506	479.84	959	3.30	42.28
Florida	8602	414.37	837	3.28	56.33
Texas	2317	361.62	800	3.27	46.94
Texas	2319	489.59	1020	3.24	51.65
Florida	11101	455.49	930	3.23	48.28
Texas	2318	478.89	1026	3.16	47.40
Texas	4504	366.62	751	3.16	44.47
Texas	4620	415.69	843	3.16	50.00
Florida	7105	602.29	1193	3.13	47.87
Texas	4622	427.88	867	3.11	46.30
Florida	8302	434.31	884	3.10	48.64
Florida	9510	430.54	866	3.10	55.08
Florida	8614	489.36	1067	3.09	57.61
Florida	1103	556.45	1130	3.08	56.03
Florida	8617	430.24	938	3.08	52.75
Florida	1102	440.72	943	3.07	52.33
Texas	2512	474.22	951	3.07	51.08
Florida	1112	410.92	862	3.04	55.24
Florida	1108	720.44	1339	3.00	57.75
Florida	1107	402.41	901	2.98	46.50
Florida	9507	573.52	1221	2.98	56.14
Florida	8605	470.20	978	2.93	58.42
Texas	4503	383.73	803	2.90	52.04
Florida	9505	668.49	1266	2.89	54.01
Florida	9908	551.69	1169	2.86	50.40
Texas	6802	345.22	703	2.85	40.81

* Male: the weighted male count in the PUMS data; Size: the sample size in the PUMS data; Margin: the margin of error for direct male proportion estimate; Twitter: the Twitter direct male proportion estimate.
** The table is in descending order of the column Margin.

tic regression coefficients $\boldsymbol{\beta}$ and the parameter r are unknown hyper-parameters. From the model, it is clear how the PUMS data and the Twitter male proportion are linked. The quantity of interest is the random effect p_i and the distribution of interest is the posterior distribution of $p_i|data$. Derived from model (5.6) and (5.7), the posterior distribution of the random effect p_i conditional on $\boldsymbol{\beta}$ and r is[25]:

$$p_i|\boldsymbol{\beta}, r, \mathbf{y} \sim Beta(n_i\bar{y}_i + rp_i^E, n_i(1-\bar{y}_i) + r(1-p_i^E)) \ , \tag{5.10}$$

where $\bar{y}_i = \hat{y}_i/n_i$. The mean and variance of the conditional posterior distribution are given by[25]:

$$p_i^* = E(p_i|\boldsymbol{\beta}, r, \mathbf{y}) = (1-B_i)\bar{y}_i + B_i p_i^E, \tag{5.11}$$

$$Var(p_i|\boldsymbol{\beta}, r, \mathbf{y}) = \frac{p_i^*(1-p_i^*)}{r+n_i+1} \ , \tag{5.12}$$

where $B_i = r/(r+n_i)$ is the shrinkage factor. The goal is to approximate the posterior distribution of $p_i|data$.

Note that the counts in Table 5.1 are non-integers. Thus the binomial model (5.6) is not appropriate. In our application we actually applied an approximate binomial likelihood following Ghitza and Gelman (2013)[16]. For an evaluation of such approximation, readers are referred to Janicki and Malec (2014)[20]. There are some other alternative approaches considered in the literature modeling survey-weighted

counts or proportions. Liu, Lahiri and Kalton (2007)[29], following up on a general recommendation of Jiang and Lahiri (2006)[21], offered a few alternative approaches to model survey-weighted proportions. Ha (2013) proposed an alternative approach for using binomial distribution that involved rounding off both the effective sample sizes and survey-weighted counts[18].

In this example, we use the R package Rgbp which implements the ADM for the Gaussian, binomial and Poisson hierarchical models to deal with the non-integer counts in the binomial-beta-logit model. The package can return estimates and their standard errors instantly without the burn-in period required by MCMC for Bayes models.

5.2.4 Results

The results from Rgbp are reported in Table 5.2, which is sorted in descending order of the margins of error computed using the Census Bureau method. The column y is the weighted number of male $\hat{y}_i$; n is the sample size n_i; $\bar{y}$ is the observed male proportion $\hat{y}_i/n_i$ from PUMS data; x_1 is the Twitter male proportion x_{i1}; $\hat{p}^E$ is the synthetic estimate of p_i, which is an approximation to $E(p_i^E|data)$; $\hat{B}$ is our approximation to $E(B|data)$; $\hat{p}$ is equal to $E(p_i|data)$; $\sigma_{\hat{p}}$ is the standard error estimate of the random effect and the last two columns are the margins of error computed by Census Bureau method and the binomial-beta-logit

model, respectively. As seen from the table, the shrinkage (Column $\hat{B}$) decreases when the sample size (Column n) in the PUMS data increases. That is, when sample size increases, the model lets the PUMS data explain more; otherwise, the posterior shrinks toward the information contained in the Twitter estimates. The margins of error generated by the binomial-beta-logit model are the standard errors (Column $\sigma_{\hat{p}}$) multiplied by 1.96 ($z_{0.025}$) and are presented in the last column (Column BB) of the table. Compared to the results computed by the Census Bureau method (Column ACS), all the margins of error are reduced and below the cutoff 3%.

5.2.5 Discussion

The contributions of this work can be summarized as :

1. We apply proper statistical models to link social media data or other big data sets with survey data and, as an example, we suggest an alternative to the method documented by the Census Bureau to estimate gender distributions at the small area level. The proposed method successfully reduces the margins of error for the male proportion estimates in PUMAs with big margins. Social media data is easy and relatively cheap to obtain compared with survey data. With more and more researchers getting involved in the research of the social media population, the quality of esti-

Table 5.2: Male Proportion Data and Results Using Binomial-Beta Model

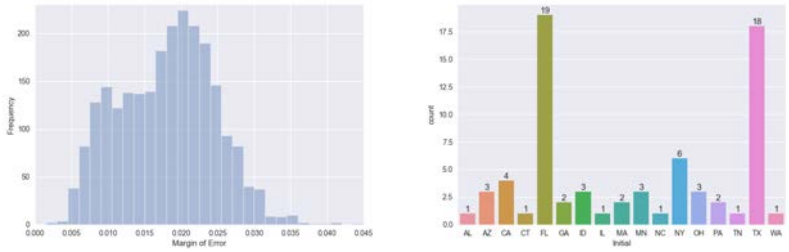
State	**Puma**	**y**	**n**	$\bar{y}$	x_1	$\hat{p}^E$	$\hat{B}$	$\hat{p}$	$\sigma_{\hat{p}}$	ACS (%)	BB (%)
Florida	9501	562.61	1164	0.483	0.624	0.486	0.533	0.485	0.0102	3.39	2.00
Florida	8604	497.51	972	0.512	0.556	0.489	0.577	0.499	0.0105	3.38	2.06
Florida	1114	527.33	1093	0.482	0.552	0.490	0.548	0.486	0.0102	3.38	2.00
Texas	2312	624.13	1172	0.533	0.484	0.493	0.531	0.512	0.0102	3.36	1.99
Texas	2506	479.84	959	0.500	0.423	0.496	0.581	0.498	0.0105	3.30	2.06
Florida	8602	414.37	837	0.495	0.563	0.489	0.613	0.491	0.0108	3.28	2.12
Texas	2317	361.62	800	0.452	0.469	0.494	0.624	0.478	0.0110	3.27	2.15
Texas	2319	489.59	1020	0.480	0.517	0.491	0.566	0.486	0.0104	3.24	2.03
Florida	11101	455.49	930	0.490	0.483	0.493	0.588	0.492	0.0106	3.23	2.07
Texas	2318	478.89	1026	0.467	0.474	0.494	0.564	0.482	0.0104	3.16	2.03
Texas	4504	366.62	751	0.488	0.445	0.495	0.639	0.493	0.0110	3.16	2.16
Texas	4620	415.69	843	0.493	0.500	0.492	0.612	0.493	0.0108	3.16	2.11
Florida	7105	602.29	1193	0.505	0.479	0.493	0.527	0.499	0.0100	3.13	1.96
Texas	4622	427.88	867	0.494	0.463	0.494	0.605	0.494	0.0107	3.11	2.10
Florida	8302	434.31	884	0.491	0.486	0.493	0.600	0.492	0.0107	3.10	2.09
Florida	9510	430.54	866	0.497	0.551	0.490	0.605	0.493	0.0107	3.10	2.11
Florida	8614	489.36	1067	0.459	0.576	0.488	0.554	0.475	0.0104	3.09	2.03
Florida	1103	556.45	1130	0.492	0.560	0.489	0.540	0.491	0.0102	3.08	1.99
Florida	8617	430.24	938	0.459	0.528	0.491	0.586	0.477	0.0106	3.08	2.08
Florida	1102	440.72	943	0.467	0.523	0.491	0.585	0.481	0.0106	3.07	2.07
Texas	2512	474.22	951	0.499	0.511	0.492	0.583	0.495	0.0105	3.07	2.06
Florida	1112	410.92	862	0.477	0.552	0.490	0.606	0.484	0.0108	3.04	2.11
Florida	1108	720.44	1339	0.538	0.577	0.488	0.498	0.513	0.0100	3.00	1.96
Florida	1107	402.41	901	0.447	0.465	0.494	0.596	0.475	0.0108	2.98	2.11
Florida	9507	573.52	1221	0.470	0.561	0.489	0.521	0.480	0.0100	2.98	1.96
Florida	8605	470.20	978	0.481	0.584	0.488	0.576	0.485	0.0105	2.93	2.06
Texas	4503	383.73	803	0.478	0.520	0.491	0.623	0.486	0.0109	2.90	2.13
Florida	9505	668.49	1266	0.528	0.540	0.490	0.512	0.509	0.0100	2.89	1.96
Florida	9908	551.69	1169	0.472	0.504	0.492	0.532	0.483	0.0101	2.86	1.97
Texas	6802	345.22	703	0.491	0.408	0.497	0.654	0.495	0.0112	2.85	2.20

mates from social media data or other useful big data resources is continuously improving. These estimates have a good prospect of serving as reliable auxiliary variables to the survey data. Introducing big data resources in data analysis promises to reduce survey cost and to produce better estimates in the future.

2. We use the existing Rgbp package, which adopts the ADM to carry out the analysis and provide fast and reliable estimates of the parameters of interest and their associated margins of error without using sampling-based methods. This proposed method can be extended to other scenarios. Since the Rgbp package can deal with normal-normal, Poisson-gamma and binomial-beta models, a wide range of continuous and count data can be handled using this package.

3. The proposed method can be applied to other geographic granularities (e.g., county and city) other than PUMA, if necessary and if proper auxiliary variables are available.

There is a limitation in this work. Since at the time of reporting the results, the one-year 2017 PUMS data was not published, we used the 2016 one-year PUMS data and 2017 Twitter data to illustrate the method. To check whether the ACS data is comparable from one year to another, we plot the histogram of the margins of error computed by the Census Bureau method for the 2015 one-year PUMS data and also



(a) *(b)*

Figure 5.2: Margins of error for the male proportion estimates (year: 2015). (a) Histogram of the estimates of the margins of error of direct estimates for the 2,378 PUMAs; (b) Number of PUMAs with margin of error greater than 3% for each state.

count the number of large-margin PUMAs in each state. There is a similar pattern in the two histograms for the margins of error between the years 2015 and 2016. There are some PUMAs with margins of error greater than the cutoff value 3%. For both of these two years, Texas and Florida have the most PUMAs with margins of error greater than the cutoff 3%. However, we can always test the method on the 2017 one-year PUMS data when it is available.

5.3 Small Area Race Distribution

5.3.1 Introduction

The data analysis example in this section is to apply the multinomial-Dirichlet-logit model to estimate the Twitter small area race proportions. The multinomial hierarchical model is implemented using the ADM we developed in this dissertation. The data used in this exam-

ple is the sample Twitter race counts for 570 PUMAs in three states (California, Florida and Texas) in the United States. A common problem in small area estimation is that the sample sizes for some areas are too small to be trusted. To solve this problem, a statistical model can be built to allow these areas to borrow strength from the entire dataset to obtain better estimates. In this data example, the small area proportion estimates are the weighted sums of the direct proportion estimates and the synthetic estimates obtained from regression on the complete covariate matrix. The model can be extended to other geographical granularities (e.g., county and city) with no limitation to PUMA. A common problem in survey datasets is non-integer weighted counts, which prevents integer-based models such as binomial, Poisson and multinomial models from being applied. For small areas, it is not proper to round the non-integer counts to the nearest integers since the decimal parts are non-trivial. Our proposed method in this data analysis example can handle this non-integer count problem in the multinomial model.

There are five races in our dataset, including Caucasian (non-Hispanic), African-American, Asian or Pacific Islander, Hispanic and Other. Section 5.3.2 provides a description of the data, including the data collection and the information extraction procedures and also briefly introduces the model for data analysis. Section 5.3.3 summarizes and

discusses the Twitter small area race proportion estimates generated by the multinomial-Dirichlet-logit model.

5.3.2 Data and Methods

In this example, we use the same Twitter data used in the example in described Section 5.2. To obtain the race counts in each PUMA, we must assign the race probabilities and the PUMA to each Twitter user based on the information in the tweets.

Individual Race Distribution: We first extract the self-reported name from each tweet and use the last name to infer the race distribution of the Twitter user. The United States Census Bureau provides a list of all surnames appearing 100 or more times in the 2010 Census[2]. There are 162,253 last names in the list, with race probabilities associated with each last name. For example, the surname 'Taylor' is 65.38% Caucasian (non-Hispanic), 28.42% African-American, 0.56% Asian or Pacific Islander, 2.46% Hispanic, and 3.18% other races. About 55.37% of the Twitter users in our sample give self-reported last names, which match last names in the list; that is, the race probabilities for 55.37% of the Twitter users in our sample can be detected. The race probabilities for the complete sample is 65.06% Caucasian (non-Hispanic), 12.06% African-American, 6.61% Asian or Pacific Islander, 13.04% Hispanic, and 2.32% other races.

Location: Location information can be inferred from the geotags contained in each tweet. A twitter user in our data can post multiple times from different locations, which results in a set of geotags for a single user. To pair a user with only one location, we first aggregate all the geotags for a single Twitter user and then associate the most frequent one with the user. Based on our data, 3,521,887 Twitter users, that is, about 96% users in the sample are geo-tagged. For illustrative purpose and low computational cost, we only select users from three states (California, Texas and Florida) based on the state information in the geotag. The reason we choose these three states is that California was carried by the Democrats while Texas was by the Republicans in all four elections from 2004 to 2016 and we saw an even balance by the two parties in Florida in these four elections. It is interesting to know how the Twitter race proportions vary with the predominant party. We feed the geotag to Bing Maps API[34] to get the longitude and latitude of each user. There are 1,173,178 sample points in the three states that were decoded successfully by the API. Then we use the PUMA shapefile[53] (e.g., by checking whether the point longitude and latitude are inside the boundary of a PUMA) to assign a PUMA to each Twitter user. There are 456,157 Twitter users successfully mapped to be inside the PUMAs in California, 312,986 in Texas and 216,462 in Florida. There are 570 PUMAs in the three states.

Race Counts: Once we have the set of the Twitter users in each PUMA and the race distribution for each user in the list, we sum up the probabilities for all the users in each PUMA for each of the five race categories. In this way, we get the non-integer counts for the five race categories in each PUMA. Let s_i denote the set of samples in i^{th} PUMA and d_{jk} denote the probability of race k, $k = 1, \ldots, K$, in the race distribution for the j^{th} Twitter user in the sample s_i. Then the count for race k in the i^{th} PUMA, denoted by y_{ik}, is calculated as follows

$$y_{ik} = \sum_{j \in s_i} d_{jk}. \tag{5.13}$$

To deal with non-integer counts for categories in our hierarchical modeling framework, we extend the binomial likelihood approach of Ghitza and Gelman (2013) to multinomial likelihood[16]. We then directly feed in the non-integer counts y_{ik} to the proposed ADM for the multinomial-Dirichlet-logit model as in Chapter 3 to obtain point and interval estimates of the small area proportions.

5.3.3 Results and Discussion

Although the model runs on the complete dataset with 570 PUMAs, results for 50 out of the 570 PUMAs are presented in Table 5.3. In the table, the categories from 1 to 5 are White (non-Hispanic), Black, Asian, Hispanic and Other, respectively. We consider the state identi-

fier as our covariate. State (Column x_{i1}) 1 is California, 0 is Florida and -1 is Texas. We observe from the table that the non-integer counts $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{i5})$, $i = 1, \ldots, 50$, do not affect the functioning of the program. Keeping the data in its original state is important under certain circumstances. Take the observation 34 in Table 5.3 as the example. This is the Leon County (Outer) PUMA in Florida, which has the minimum group size in the sample in this table. If a multinomial model is applied using MCMC or some other approaches which require integer input, the rounded counts for this group will be $(0, 0, 0, 1, 0)$ and the observed proportion is $\bar{y} = (0, 0, 0, 1, 0)$ after rounding, which is significantly different from the original proportion $(0.17, 0.04, 0.01, 0.78, 0.00)$. Thus rounding will affect the accuracy of estimates when the sample size is small. Our proposed procedure solves this problem and does not require integer input for the multinomial model. To solve the problem of unreliability in direct estimates for PUMAs with small sample sizes, our procedure estimates individual shrinkage B_i, $i = 1, \ldots, 50$, for each PUMA. As observed in Table 5.3, the shrinkage $\hat{B}_i$ increases as the group size $n_i = \sum_{i=1}^{5} y_{ik}$ decreases. That is, when group size is small in a PUMA, the estimate $\hat{\boldsymbol{p}}_i$ will shrink toward the state mean $\hat{\boldsymbol{p}}_i^E$, which is determined by the logistic regression. The Leon County (Outer) PUMA in Florida has the largest shrinkage in the sample, with $\hat{\boldsymbol{p}}_i$ closer to $\hat{\boldsymbol{p}}_i^E$. The results from the model indicates that the Twitter White and Black percentages are the lowest in California and the high-

est in Texas while Asian and Hispanic percentages are the highest in California and the lowest in Texas. In this example, we only have state identifier as the covariate; however, it is possible to introduce covariates other than the state identifier in the multinomial-Dirichlet-logit model if the readers are interested in other factors that can affect the small area race proportions and if such data is available.

Table 5.3: Twitter Race Count Data and Analysis Results Using ADM

	Data							$\hat{\boldsymbol{\beta}} = \begin{pmatrix} 2.673 & -0.036 \\ 1.129 & -0.107 \\ 0.639 & 0.176 \\ 1.588 & 0.060 \end{pmatrix}, \hat{r} = 59.420$										
obs i	n_i	y_{i1}	y_{i2}	y_{i3}	y_{i4}	y_{i5}	x_{i1}	$\hat{p}_{i1}^E$	$\hat{p}_{i2}^E$	$\hat{p}_{i3}^E$	$\hat{p}_{i4}^E$	$\hat{p}_{i5}^E$	$\hat{B}_i$	$\hat{p}_{i1}$	$\hat{p}_{i2}$	$\hat{p}_{i3}$	$\hat{p}_{i4}$	$\hat{p}_{i5}$
1	1583	959.66	173.41	116.95	282.56	50.41	1	0.554	0.110	0.090	0.206	0.040	0.036	0.604 (0.012)	0.110 (0.008)	0.074 (0.006)	0.179 (0.009)	0.032 (0.004)
2	1042	519.25	110.55	113.00	264.66	34.55	1	0.554	0.110	0.090	0.206	0.040	0.054	0.501 (0.015)	0.106 (0.009)	0.107 (0.009)	0.251 (0.013)	0.034 (0.005)
3	874	544.68	97.14	54.02	149.78	28.39	1	0.554	0.110	0.090	0.206	0.040	0.064	0.619 (0.016)	0.111 (0.010)	0.064 (0.008)	0.174 (0.012)	0.033 (0.006)
4	729	457.09	80.64	47.79	118.47	25.00	1	0.554	0.110	0.090	0.206	0.040	0.075	0.622 (0.017)	0.111 (0.011)	0.067 (0.009)	0.166 (0.013)	0.035 (0.007)
5	718	329.79	72.82	86.56	204.89	23.95	1	0.554	0.110	0.090	0.206	0.040	0.076	0.467 (0.018)	0.102 (0.011)	0.118 (0.012)	0.279 (0.016)	0.034 (0.006)
6	704	394.23	70.79	43.10	175.61	20.27	1	0.554	0.110	0.090	0.206	0.040	0.078	0.560 (0.018)	0.101 (0.011)	0.063 (0.009)	0.246 (0.016)	0.030 (0.006)
7	689	457.87	81.35	44.05	84.11	21.61	1	0.554	0.110	0.090	0.206	0.040	0.079	0.656 (0.017)	0.117 (0.012)	0.066 (0.009)	0.129 (0.012)	0.032 (0.006)
8	553	294.14	61.71	42.72	137.05	17.38	1	0.554	0.110	0.090	0.206	0.040	0.097	0.534 (0.020)	0.111 (0.013)	0.078 (0.011)	0.244 (0.017)	0.032 (0.007)
9	400	154.07	31.54	29.05	174.81	10.53	1	0.554	0.110	0.090	0.206	0.040	0.129	0.407 (0.023)	0.083 (0.013)	0.075 (0.012)	0.407 (0.023)	0.028 (0.008)
10	380	146.03	36.44	79.91	102.39	15.23	1	0.554	0.110	0.090	0.206	0.040	0.135	0.407 (0.023)	0.098 (0.014)	0.194 (0.019)	0.261 (0.021)	0.040 (0.009)
11	365	188.59	35.31	22.87	104.91	13.32	1	0.554	0.110	0.090	0.206	0.040	0.140	0.522 (0.024)	0.099 (0.014)	0.066 (0.012)	0.276 (0.022)	0.037 (0.009)
12	350	166.93	36.25	21.13	115.15	10.54	1	0.554	0.110	0.090	0.206	0.040	0.145	0.488 (0.025)	0.105 (0.015)	0.065 (0.012)	0.311 (0.023)	0.032 (0.009)
13	289	115.12	21.66	25.85	117.10	9.28	1	0.554	0.110	0.090	0.206	0.040	0.171	0.425 (0.027)	0.015 (0.015)	0.089 (0.015)	0.371 (0.026)	0.033 (0.010)
14	272	159.22	31.44	11.91	60.62	8.81	1	0.554	0.110	0.090	0.206	0.040	0.179	0.580 (0.027)	0.115 (0.017)	0.052 (0.012)	0.220 (0.023)	0.034 (0.010)
15	217	113.71	19.19	23.05	54.24	6.81	1	0.554	0.110	0.090	0.206	0.040	0.215	0.531 (0.030)	0.093 (0.017)	0.103 (0.018)	0.241 (0.026)	0.033 (0.011)
16	35	22.33	4.20	6.12	1.32	1.04	1	0.554	0.110	0.090	0.206	0.040	0.629	0.585 (0.051)	0.114 (0.033)	0.121 (0.034)	0.144 (0.039)	0.036 (0.019)
17	6	2.62	0.27	0.95	2.01	0.15	1	0.554	0.110	0.090	0.206	0.040	0.908	0.544 (0.064)	0.104 (0.039)	0.096 (0.038)	0.218 (0.055)	0.038 (0.024)
18	6840	3137.18	605.84	470.24	2436.03	190.71	0	0.571	0.122	0.075	0.193	0.039	0.009	0.460 (0.006)	0.089 (0.003)	0.069 (0.003)	0.355 (0.006)	0.028 (0.002)
19	4429	2830.48	554.18	276.12	625.07	143.14	0	0.571	0.122	0.075	0.193	0.039	0.013	0.638 (0.007)	0.125 (0.005)	0.063 (0.004)	0.142 (0.005)	0.032 (0.003)
20	3148	1643.04	297.33	245.14	866.82	95.67	0	0.571	0.122	0.075	0.193	0.039	0.019	0.523 (0.009)	0.095 (0.005)	0.078 (0.005)	0.274 (0.008)	0.031 (0.003)
21	1270	777.56	162.97	75.85	214.16	39.46	0	0.571	0.122	0.075	0.193	0.039	0.045	0.610 (0.013)	0.128 (0.009)	0.060 (0.007)	0.170 (0.010)	0.031 (0.005)
22	828	297.42	59.93	51.41	396.79	22.46	0	0.571	0.122	0.075	0.193	0.039	0.067	0.373 (0.016)	0.076 (0.009)	0.063 (0.008)	0.460 (0.017)	0.028 (0.006)
23	691	356.14	74.14	41.75	197.46	21.52	0	0.571	0.122	0.075	0.193	0.039	0.079	0.520 (0.018)	0.108 (0.011)	0.062 (0.009)	0.278 (0.016)	0.032 (0.006)
24	643	403.80	70.17	43.77	104.86	20.39	0	0.571	0.122	0.075	0.193	0.039	0.085	0.623 (0.018)	0.110 (0.012)	0.069 (0.010)	0.166 (0.014)	0.032 (0.007)
25	560	206.46	40.34	44.39	251.70	17.11	0	0.571	0.122	0.075	0.193	0.039	0.096	0.388 (0.020)	0.077 (0.011)	0.079 (0.011)	0.425 (0.020)	0.031 (0.007)
26	529	377.12	61.03	27.72	44.93	18.19	0	0.571	0.122	0.075	0.193	0.039	0.101	0.699 (0.019)	0.116 (0.013)	0.055 (0.009)	0.096 (0.012)	0.035 (0.008)
27	347	166.72	28.39	21.99	121.38	8.52	0	0.571	0.122	0.075	0.193	0.039	0.146	0.494 (0.025)	0.088 (0.014)	0.065 (0.012)	0.327 (0.023)	0.027 (0.008)
28	256	160.92	34.00	13.21	39.93	7.94	0	0.571	0.122	0.075	0.193	0.039	0.188	0.618 (0.027)	0.131 (0.019)	0.056 (0.013)	0.163 (0.021)	0.033 (0.010)
29	239	170.06	35.46	12.52	13.23	7.73	0	0.571	0.122	0.075	0.193	0.039	0.199	0.684 (0.027)	0.143 (0.020)	0.057 (0.013)	0.083 (0.016)	0.034 (0.010)
30	203	136.32	22.54	9.11	28.64	6.40	0	0.571	0.122	0.075	0.193	0.039	0.226	0.649 (0.030)	0.114 (0.020)	0.052 (0.014)	0.153 (0.022)	0.033 (0.011)
31	139	91.23	12.23	5.97	26.12	3.46	0	0.571	0.122	0.075	0.193	0.039	0.299	0.631 (0.034)	0.098 (0.021)	0.052 (0.016)	0.189 (0.028)	0.029 (0.012)
32	100	39.86	5.76	5.23	46.30	2.86	0	0.571	0.122	0.075	0.193	0.039	0.373	0.463 (0.040)	0.082 (0.022)	0.061 (0.019)	0.362 (0.040)	0.033 (0.014)
33	5	3.68	0.16	0.06	0.99	0.11	0	0.571	0.122	0.075	0.193	0.039	0.922	0.067 (0.030)	0.115 (0.042)	0.070 (0.033)	0.193 (0.049)	0.038 (0.024)
34	1	0.17	0.04	0.01	0.77	0.00	0	0.571	0.122	0.075	0.193	0.039	0.983	0.564 (0.095)	0.121 (0.045)	0.074 (0.035)	0.203 (0.116)	0.039 (0.026)
35	2173	1363.53	289.14	188.84	261.40	70.10	-1	0.585	0.134	0.062	0.180	0.039	0.027	0.626 (0.010)	0.133 (0.007)	0.086 (0.006)	0.122 (0.007)	0.032 (0.004)
36	924	597.88	129.85	81.52	83.59	31.17	-1	0.585	0.134	0.062	0.180	0.039	0.060	0.643 (0.015)	0.140 (0.011)	0.087 (0.009)	0.096 (0.009)	0.034 (0.006)
37	914	582.47	137.04	44.49	119.65	30.34	-1	0.585	0.134	0.062	0.180	0.039	0.061	0.634 (0.015)	0.149 (0.011)	0.049 (0.007)	0.134 (0.011)	0.034 (0.006)
38	735	482.61	106.24	43.77	78.90	23.47	-1	0.585	0.134	0.062	0.180	0.039	0.075	0.651 (0.017)	0.144 (0.012)	0.060 (0.008)	0.113 (0.011)	0.032 (0.006)
39	541	293.81	93.08	54.47	78.78	20.86	-1	0.585	0.134	0.062	0.180	0.039	0.099	0.547 (0.020)	0.168 (0.015)	0.097 (0.012)	0.149 (0.015)	0.039 (0.008)
40	498	303.62	62.41	33.04	82.17	16.77	-1	0.585	0.134	0.062	0.180	0.039	0.107	0.607 (0.021)	0.126 (0.014)	0.066 (0.011)	0.167 (0.016)	0.034 (0.008)
41	489	310.25	65.24	22.29	76.74	14.47	-1	0.585	0.134	0.062	0.180	0.039	0.108	0.629 (0.021)	0.134 (0.015)	0.047 (0.009)	0.159 (0.016)	0.031 (0.007)
42	426	288.39	69.29	23.34	28.39	16.58	-1	0.585	0.134	0.062	0.180	0.039	0.122	0.666 (0.021)	0.159 (0.017)	0.056 (0.010)	0.080 (0.012)	0.039 (0.009)
43	272	182.33	39.24	8.90	33.45	8.09	-1	0.585	0.134	0.062	0.180	0.039	0.179	0.655 (0.026)	0.142 (0.019)	0.038 (0.011)	0.133 (0.019)	0.031 (0.010)
44	254	163.88	35.83	14.79	31.75	7.74	-1	0.585	0.134	0.062	0.180	0.039	0.190	0.634 (0.027)	0.140 (0.020)	0.059 (0.013)	0.135 (0.019)	0.032 (0.010)
45	214	142.98	27.61	13.10	23.00	7.32	-1	0.585	0.134	0.062	0.180	0.039	0.217	0.650 (0.029)	0.130 (0.020)	0.061 (0.015)	0.123 (0.020)	0.035 (0.011)
46	150	99.12	23.53	8.58	14.18	4.59	-1	0.585	0.134	0.062	0.180	0.039	0.284	0.639 (0.033)	0.150 (0.025)	0.059 (0.016)	0.119 (0.022)	0.033 (0.012)
47	67	36.84	9.64	2.99	15.83	1.70	-1	0.585	0.134	0.062	0.180	0.039	0.470	0.566 (0.044)	0.139 (0.031)	0.053 (0.020)	0.210 (0.036)	0.032 (0.016)
48	64	36.23	8.79	2.91	13.03	3.04	-1	0.585	0.134	0.062	0.180	0.039	0.481	0.575 (0.045)	0.136 (0.031)	0.053 (0.020)	0.192 (0.035)	0.043 (0.018)
49	61	37.49	8.78	3.60	9.07	2.06	-1	0.585	0.134	0.062	0.180	0.039	0.493	0.600 (0.045)	0.139 (0.032)	0.060 (0.022)	0.164 (0.034)	0.036 (0.017)
50	3	1.78	1.03	0.02	0.07	0.10	-1	0.585	0.134	0.062	0.180	0.039	0.952	0.585 (0.062)	0.144 (0.057)	0.059 (0.031)	0.172 (0.055)	0.039 (0.024)

Chapter 6: Discussion and Future Research

6.1 Discussion

Morris and his students developed the ADM for a number of univariate distributions mentioned in Chapter 2. This dissertation fills in an important research gap by extending the ADM beyond univariate distributions, specifically the important multinomial-Dirichlet-logit hierarchical distribution. We have demonstrated results from the ADM and MCMC are virtually the same. However, the computational speed of ADM is hundreds of times faster than that of MCMC. When compared to the classical EB procedures, the ADM is superior to the EB methods in terms of criteria used in the EB framework, especially for small samples. All the advantages we have observed in the ADM for random effect estimation in the multinomial-Dirichlet-logit model have also been reported by Morris and his students in their studies of the ADM for a series of hierarchical univariate Bayes models. We now discuss our plan for future research.

6.2 ADM for COM-Poisson Bayes Model

We would like to extend the ADM to the hierarchical COM-Poisson model. The COM-Poisson distribution can be used for both over-dispersed and under-dispersed data. We now present the descriptive and inferential forms of the conjugate hierarchical COM-Poisson model in Section 6.2.1 and Section 6.2.2 and discuss the questions to be solved in developing the ADM for the hierarchical COM-Poisson model.

6.2.1 The Descriptive Model

The COM-Poisson distribution has the density

$$f(y_i|\lambda_i,\nu) = \frac{\lambda_i^{y_i}}{(y_i!)^\nu} \cdot \frac{1}{Z(\lambda_i,\nu)},\ y_i = 0,1,2,\ldots, \tag{6.1}$$

where

$$Z(\lambda_i,\nu) = \sum_{j=0}^{\infty} \frac{\lambda_i^j}{(j!)^\nu}. \tag{6.2}$$

We assume that ν is known and is the same for all observations for the time being.

Level 1 (The individual model) : The count y_i given the unknown individual parameter λ_i has the following density

$$y_i|\lambda_i \sim COM - Poisson(\lambda_i,\nu), \tag{6.3}$$

with the parameter ν known.

Level 2 (The structural model) : The conjugate prior for the individual parameter λ_i given the hyper-parameters a and b is given by

$$h(\lambda_i|a,b) \propto \lambda_i^{a-1} Z^{-b}(\lambda_i,\nu), \tag{6.4}$$

with the parameter ν known.

Before deciding upon a third level hyper-prior to prevent posterior impropriety, the inferential model must be given for the purpose of obtaining the likelihood function of the hyper-parameters.

6.2.2 The Inferential Model

Level 1 (The marginal model for the observations). The observation counts y_i given the hyper-parameters a and b with the individual parameter λ_i integrated out is distributed with density

$$\begin{aligned} f(y_i|a,b) &= \int_0^\infty f(y_i|\lambda_i,\nu) h(\lambda_i|a,b) d\lambda_i \\ &= \kappa^{-1}(a,b) \int_0^\infty \frac{\lambda_i^{y_i}}{Z(\lambda_i,\nu)} \lambda_i^{a-1} Z^{-b}(\lambda_i,\nu) d\lambda_i \\ &= \kappa^{-1}(a,b) \int_0^\infty \lambda_i^{y_i+a-1} Z^{-(b+1)}(\lambda_i,\nu) d\lambda_i \\ &= \frac{\kappa(y_i+a, b+1)}{\kappa(a,b)}, \end{aligned} \tag{6.5}$$

where $\kappa(a,b) = \int_0^\infty \lambda_i^{a-1} Z^{-b}(\lambda_i,\nu) d\lambda_i$.

Level 2 (The conditional model for the individual parameters). The conditional posterior distribution for the individual parameter λ_i has the same distribution as in equation (6.4) with updated parameters

$$h(\lambda_i|a, b, data) \propto \lambda_i^{y_i+a-1} Z^{-(b+1)}(\lambda_i, \nu), \tag{6.6}$$

where the updated parameters are $y_i + a$ and $b + 1$.

From equation (6.5), the likelihood $L(a, b)$ is

$$L(a, b) = \prod_{i=1}^{N} \frac{\kappa(y_i + a, b + 1)}{\kappa(a, b)}. \tag{6.7}$$

6.2.3 Discussion

To decide upon whether a third level adjustment is necessary, the work in the next step is to prove whether there exists a sufficient condition of the observed data for the likelihood (6.7) to be proper, that is to prove whether

$$\int_0^\infty \int_0^\infty \prod_0^\infty \frac{\kappa(y_i + a, b + 1)}{\kappa(a, b)} dadb \tag{6.8}$$

converges. Direct integral is hard. The common method is to find integrable bounds for the individual equation (6.5). If the lower bound of the likelihood (6.7) is improper, a third level adjustment is needed and the form of the third level hyper-prior must be decided. Then, the questions to be solved include whether there are closed-forms or

approximate closed-forms of the mean and variance for the Level 2 distribution (6.4) in the descriptive model. This is important because the mean of this distribution can be linked with a regression on the covariates in some way. And the closed-forms of the mean and variance for the Level 2 distribution (6.6) in the inferential model are promising in introducing the individual shrinkage factors, which are convenient in explaining the results. Once the posterior propriety is guaranteed and there are closed-forms for the mean and variance in equation (6.4) and equation (6.6), it is possible to develop the ADM for this COM-Poisson distribution.

Appendices

A.1 Proof of Lemma 3.1

If the group i is interior ($d_i = K$, $n_i \geq K$, $K \geq 3$), we can derive lower and upper bound for the Dirichlet-multinomial probability mass function with respect to $\boldsymbol{\beta}$ and r as follows. All the bounds in this proof are up to a constant multiple.

$$\begin{aligned} p(\boldsymbol{y_i}|r,\boldsymbol{\beta}) &\propto \frac{\Gamma(r)}{\Gamma(n_i+r)} \prod_{k=1}^{K} \frac{\Gamma(y_{ik}+rp_{ik}^E)}{\Gamma(rp_{ik}^E)} \\ &= \frac{1}{(r+n_i-1)\cdot\ldots\cdot r} \prod_{k=1}^{K} (rp_{ik}^E + y_{ik} - 1)\cdot\ldots\cdot rp_{ik}^E \\ &= \frac{r^K \prod_{k=1}^{K} p_{ik}^E}{r(r+1)\cdot\ldots\cdot(r+K-1)} \cdot \frac{\prod_{k:y_{ik}\geq 2}(rp_{ik}^E+1)\cdot\ldots\cdot(rp_{ik}^E+y_{ik}-1)}{(r+K)\cdot\ldots\cdot(r+n_i-1)} \\ &\leq \frac{r^2}{(r+1)(r+2)} \prod_{k=1}^{K} p_{ik}^E \end{aligned} \tag{9}$$

The inequality holds considering $K \geq 3$ and $1 \leq y_{ik} \leq n_i - 1$ for k=1,..., K.

A lower bound for the Dirichlet-multinomial probability mass function is

$$
\begin{aligned}
p(\boldsymbol{y_i}|r,\boldsymbol{\beta}) &\propto \frac{\Gamma(r)}{\Gamma(n_i+r)}\prod_{k=1}^{K}\frac{\Gamma(y_{ik}+rp_{ik}^E)}{\Gamma(rp_{ik}^E)} \\
&= \frac{1}{(r+n_i-1)\cdot\ldots\cdot r}\prod_{k=1}^{K}(rp_{ik}^E+y_{ik}-1)\cdot\ldots\cdot rp_{ik}^E \\
&\geq \frac{1}{(r+n_i-1)\cdot\ldots\cdot r}\prod_{k=1}^{K}(rp_{ik}^E)^{y_{ik}} \\
&= \frac{r^{n_i}}{(r+n_i-1)\cdot\ldots\cdot r}\prod_{k=1}^{K}(p_{ik}^E)^{y_{ik}} \\
&\geq \left(\frac{r}{r+n_{max}}\right)^{n_i}\prod_{k=1}^{K}(p_{ik}^E)^{y_{ik}}
\end{aligned}
\tag{10}
$$

where $n_{max} = max\{n_1,\ldots,n_N\}$. The first inequality holds because all $y_{ik} \geq 1$ for interior group i.

Similarly, for intermediate group i ($2 \leq d_i \leq K-1$), the upper bound for the Dirichlet-multinomial probability mass function with respect to $\boldsymbol{\beta}$ and r is $\frac{r}{r+1}\prod_{k\in W_i}p_{ik}^E$, up to constant multiple, given $d_i \geq 2$.

$$
\begin{aligned}
p(\boldsymbol{y_i}|r,\boldsymbol{\beta}) &\propto \frac{\Gamma(r)}{\Gamma(n_i+r)}\prod_{k\in W_i}\frac{\Gamma(y_{ik}+rp_{ik}^E)}{\Gamma(rp_{ik}^E)} \\
&= \frac{1}{(r+n_i-1)\cdot\ldots\cdot r}\prod_{k\in W_i}(rp_{ik}^E+y_{ik}-1)\cdot\ldots\cdot rp_{ik}^E \\
&= \frac{r^{d_i}\prod_{k\in W_i}p_{ik}^E}{r(r+1)\cdot\ldots\cdot(r+d_i-1)}\cdot\frac{\prod_{k:y_{ik}\geq 2}(rp_{ik}^E+1)\cdot\ldots\cdot(rp_{ik}^E+y_{ik}-1)}{(r+d_i)\cdot\ldots\cdot(r+n_i-1)} \\
&\leq \frac{r}{r+1}\prod_{k\in W_i}p_{ik}^E
\end{aligned}
\tag{11}
$$

And a lower bound for intermediate group i is

$$
\begin{aligned}
p(\boldsymbol{y_i}|r,\boldsymbol{\beta}) &\propto \frac{\Gamma(r)}{\Gamma(n_i+r)} \prod_{k\in W_i} \frac{\Gamma(y_{ik}+rp_{ik}^E)}{\Gamma(rp_{ik}^E)} \\
&= \frac{1}{(r+n_i-1)\cdot\ldots\cdot r} \prod_{k\in W_i} (rp_{ik}^E+y_{ik}-1)\cdot\ldots\cdot rp_{ik}^E \\
&\geq \frac{1}{(r+n_i-1)\cdot\ldots\cdot r} \prod_{k\in W_i} (rp_{ik}^E)^{y_{ik}} \\
&= \frac{r^{n_i}}{(r+n_i-1)\cdot\ldots\cdot r} \prod_{k\in W_i} (p_{ik}^E)^{y_{ik}} \\
&\geq \left(\frac{r}{r+n_{max}}\right)^{n_i} \prod_{k\in W_i} (p_{ik}^E)^{y_{ik}}
\end{aligned}
\tag{12}
$$

For extreme group i ($d_i = 1$), assume the total mass n_i ($n_i \geq 1$) fall in the category j. That is, $y_{ik} = n_i$ when $k = j$ and $y_{ik} = 0$ when $k \neq j$. The upper bound for the Dirichlet-multinomial probability mass function of the group i with respect to $\boldsymbol{\beta}$ and r is

$$
\begin{aligned}
p(\boldsymbol{y_i}|r,\boldsymbol{\beta}) &\propto \frac{\Gamma(r)}{\Gamma(n_i+r)} \cdot \frac{\Gamma(y_{ij}+rp_{ij}^E)}{\Gamma(rp_{ij}^E)} = \frac{\Gamma(r)}{\Gamma(n_i+r)} \cdot \frac{\Gamma(n_i+rp_{ij}^E)}{\Gamma(rp_{ij}^E)} \\
&= \frac{(rp_{ij}^E+n_i-1)\cdot\ldots\cdot(rp_{ij}^E+1)rp_{ij}^E}{(r+n_i-1)\cdot\ldots\cdot(r+1)r} \\
&= p_{ij}^E \cdot \frac{(rp_{ij}^E+n_i-1)\cdot\ldots\cdot(rp_{ij}^E+1)}{(r+n_i-1)\cdot\ldots\cdot(r+1)} \\
&= p_{ij}^E \sum_{s=1}^{n_i-1} \frac{rp_{ij}^E+s}{r+s} \leq p_{ij}^E
\end{aligned}
\tag{13}
$$

The inequality holds because the ratios $\frac{rp_{ij}^E+s}{r+s}$, $s = 1,\ldots,n_i-1$, in the second

term are less or equal to 1. A lower bound for extreme group i is

$$\begin{aligned} p(\boldsymbol{y_i}|r,\boldsymbol{\beta}) &\propto \frac{\Gamma(r)}{\Gamma(n_i+r)} \cdot \frac{\Gamma(y_{ij}+rp_{ij}^E)}{\Gamma(rp_{ij}^E)} = \frac{\Gamma(r)}{\Gamma(n_i+r)} \cdot \frac{\Gamma(n_i+rp_{ij}^E)}{\Gamma(rp_{ij}^E)} \\ &= \frac{(rp_{ij}^E+n_i-1)\cdot\ldots\cdot(rp_{ij}^E+1)rp_{ij}^E}{(r+n_i-1)\cdot\ldots\cdot(r+1)r} \\ &= \frac{rp_{ij}^E+n_i-1}{r+n_i-1}\cdot\ldots\cdot\frac{rp_{ij}^E}{r} \\ &\geq (p_{ij}^E)^{n_i} \end{aligned} \tag{14}$$

The inequality holds because each ratio in the product is greater than or equal to p_{ij}^E.

A.2 Proof of Lemma 3.2

With no intermediate or extreme groups in the data, an upper bound for the likelihood function $L(\boldsymbol{\beta}, r)$ is the product of the upper bound for the individual Dirichlet-multinomial probability mass functions as in equation (9):

$$\prod_{i=1}^{N}\Big(\frac{r^2}{(r+1)(r+2)}\prod_{k=1}^{K}p_{ik}^E\Big) = \frac{r^{2N}}{(r+1)^N(r+2)^N}\prod_{i=1}^{N}\prod_{k=1}^{K}p_{ik}^E \tag{15}$$

The upper bound factors into a function of $\boldsymbol{\beta}$ and a function of r. And a lower bound for the likelihood function $L(\boldsymbol{\beta}, r)$ is the product of the lower bound for the individual Dirichlet-multinomial probability mass functions as in equation

(10):

$$\prod_{i=1}^{N}\Big(\frac{r}{r+n_{max}}\Big)^{n_i}\prod_{k=1}^{K}(p_{ik}^{E})^{y_{ik}}=\Big(\frac{r}{r+n_{max}}\Big)^{\sum_{i=1}^{N}n_i}\prod_{i=1}^{N}\prod_{k=1}^{K}(p_{ik}^{E})^{y_{ik}} \tag{16}$$

A.3 Proof of Theorem 3.1

The posterior density $p(\boldsymbol{\beta},r)=L(\boldsymbol{\beta},r)/r^2$ is bounded from above, up to a multiple constant

$$p(\boldsymbol{\beta},r)\leq\frac{r^{2N-2}}{(r+1)^N(r+2)^N}\prod_{i=1}^{N}\prod_{k=1}^{K}p_{ik}^{E} \tag{17}$$

The posterior density is proper if

$$\begin{aligned}&\int_{\mathbb{R}^{(K-1)\times q}}\int_0^\infty\frac{r^{2N-2}}{(r+1)^N(r+2)^N}\prod_{i=1}^{N}\prod_{k=1}^{K}p_{ik}^{E}drd\boldsymbol{\beta}\\=&\int_0^\infty\frac{r^{2N-2}}{(r+1)^N(r+2)^N}dr\int_{\mathbb{R}^{(K-1)\times q}}\prod_{i=1}^{N}\prod_{k=1}^{K}p_{ik}^{E}d\boldsymbol{\beta}<\infty\end{aligned} \tag{18}$$

First consider the integral with respect to r in the upper bound. When $N\geq 1$,

$$\begin{aligned}\int_0^\infty\frac{r^{2N-2}}{(r+1)^N(r+2)^N}dr&=\int_0^\infty\frac{r^{2N-2}}{(r+1)^{N-1}(r+2)^{N-1}}\frac{1}{(r+1)(r+2)}dr\\&\leq\int_0^\infty\frac{1}{(r+1)(r+2)}dr=log(2)\end{aligned} \tag{19}$$

Then consider the integral with respect to $\boldsymbol{\beta}$. Choose q sub-groups, whose index set is denoted by W_{sub}, such that the $q\times q$ covariate matrix of the

sub-groups is still of full rank q.

$$\prod_{i=1}^{N}\prod_{k=1}^{K} p_{ik}^{E} \leq \prod_{i\in W_{sub}}\prod_{k=1}^{K} p_{ik}^{E} = \prod_{i\in W_{sub}} \frac{\prod_{k=1}^{K-1} e^{\boldsymbol{x}_i'\boldsymbol{\beta}_k}}{(1+\sum_{j=1}^{K-1} e^{\boldsymbol{x}_i'\boldsymbol{\beta}_j})^K} \tag{20}$$

The integration of this upper bound with respect to $\boldsymbol{\beta}$ factors into $(K-1)\times q$ separate integrations after linear transformations, $h_{ik} = \boldsymbol{x}_i'\boldsymbol{\beta}_k$ for all $i \in W_{sub}$, whose Jacobian is a constant:

$$\begin{aligned} &\int_{\mathbb{R}^{(K-1)\times q}} \prod_{i\in W_{sub}} \frac{\prod_{k=1}^{K-1} e^{\boldsymbol{x}_i'\boldsymbol{\beta}_k}}{(1+\sum_{j=1}^{K-1} e^{\boldsymbol{x}_i'\boldsymbol{\beta}_j})^K} d\boldsymbol{\beta} \\ \propto &\prod_{i\in W_{sub}} \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \frac{\prod_{k=1}^{K-1} e^{h_{ik}}}{(1+\sum_{j=1}^{K-1} e^{h_{ij}})^K} dh_{i,1}\ldots dh_{i,K-1} = (\frac{1}{(K-1)!})^N. \end{aligned} \tag{21}$$

A.4 Proof of Corollary 3.1

Regarding the sufficient conditions for posterior propriety, an upper bound for $L(\boldsymbol{\beta}, r)$ up to a constant multiple is

$$\begin{aligned} L(\boldsymbol{\beta}, r) &\propto \prod_{i=1}^{N} \frac{\Gamma(r)}{\Gamma(n_i+r)} \prod_{k=1}^{K} \frac{\Gamma(y_{ik}+rp_{ik}^E)}{\Gamma(rp_{ik}^E)} < \prod_{i\in W_y} \frac{\Gamma(r)}{\Gamma(n_i+r)} \prod_{k=1}^{K} \frac{\Gamma(y_{ik}+rp_{ik}^E)}{\Gamma(rp_{ik}^E)} \\ &\leq \prod_{i\in W_y} \frac{r^2}{(r+1)(r+2)} \prod_{k=1}^{K} p_{ik}^E = \frac{r^{2N_y}}{(r+1)^{N_y}(r+2)^{N_y}} \prod_{i\in W_y}\prod_{k=1}^{K} p_{ik}^E \end{aligned} \tag{22}$$

The first inequality holds because both the upper bound for intermediate group i, $\frac{r}{r+1}\prod_{k\in W_i} p_{ik}^E$, and the upper bound for extreme group i, p_{ij}^E, are less than 1.

The upper bound for $L(\boldsymbol{\beta}, r)$ in equation(22) is the same as the upper bound

if all the intermediate and extreme groups are removed from the data. Therefore, the sufficient condition can be determined by the interiors groups in the data.

A.5 Bounds with Non-integer Counts

Since we hope to apply the ADM for the multinomial-Dirichlet-logit model to non-integer counts, we also prove the upper bound for the Dirichlet-multinomial function for groups with non-integer counts. Before proving the bounds, we will first prove a lemma.

Lemma A.1 *For any $y > 0$, there exists positive constants $0 < c_1 \leq c_2 < \infty$, which only depend on y such that*

$$c_1 r(r+1)^{y-1} \leq \frac{\Gamma(r+y)}{\Gamma(r)} \leq c_2 r(r+1)^{y-1}, \tag{23}$$

for any $r > 0$.

Proof: Let

$$\begin{aligned} g(r) &= \frac{\Gamma(r+y)}{r\Gamma(r)(r+1)^{y-1}} \\ &= \frac{\Gamma(r+y)}{\Gamma(r+1)(r+1)^{y-1}}. \end{aligned} \tag{24}$$

From this expression, $g(r)$ is well-defined and continuous on $(0, \infty)$. Since

$$\lim_{x\to\infty} \frac{\Gamma(x+\alpha)}{\Gamma(x)x^{\alpha}} = 1, \tag{25}$$

for any $\alpha \in \mathbb{C}$;

$$\lim_{r\to\infty} g(r) = \lim_{r\to\infty} \frac{\Gamma(r+y)}{\Gamma(r+1)(r+1)^{y-1}} = 1, \tag{26}$$

for any $y > 0$. And $g(r) > 0$ for any $r \in (0, \infty)$. Then, $g(r)$ is bounded from both below and above in $(0, \infty)$. That is,

$$c_1 \leq g(r) \leq c_2; \tag{27}$$

thus,

$$c_1 r(r+1)^{y-1} \leq \frac{\Gamma(r+y)}{\Gamma(r)} \leq c_2 r(r+1)^{y-1}. \tag{28}$$

Use this lemma, for interior group i ($d_i = K$, $n_i \geq K$, $K \geq 3$):

$$\begin{aligned} p(\boldsymbol{y_i}|r, \boldsymbol{\beta}) &\propto \frac{\Gamma(r)}{\Gamma(n_i + r)} \prod_{k=1}^{K} \frac{\Gamma(y_{ik} + rp_{ik}^E)}{\Gamma(rp_{ik}^E)} \\ &\leq \frac{1}{(r+n_i-1)\cdot \ldots \cdot r} \prod_{k=1}^{K} rp_{ik}^E (rp_{ik}^E + 1)^{y_{ik}-1} \\ &\leq \frac{1}{(r+n_i-1)\cdot \ldots \cdot r} \prod_{k=1}^{K} r(r+1)^{y_{ik}-1} (p_{ik}^E)^{min(y_{ik},1)} \\ &= \frac{r^K (r+1)^{n_i-K}}{(r+n_i-1)\cdot \ldots \cdot r} \prod_{k=1}^{K} (p_{ik}^E)^{min(y_{ik},1)} \\ &\leq \frac{r^2}{(r+1)(r+2)} \prod_{k=1}^{K} (p_{ik}^E)^{min(y_{ik},1)}, \end{aligned} \tag{29}$$

up to a constant.

Similarly, for intermediate group i ($2 \leq d_i \leq K-1$):

$$
\begin{aligned}
p(\boldsymbol{y_i}|r,\boldsymbol{\beta}) &\propto \frac{\Gamma(r)}{\Gamma(n_i+r)} \prod_{k\in W_i} \frac{\Gamma(y_{ik}+rp_{ik}^E)}{\Gamma(rp_{ik}^E)} \\
&\leq \frac{1}{(r+n_i-1)\cdot\ldots\cdot r} \prod_{k\in W_i} rp_{ik}^E(rp_{ik}^E+1)^{y_{ik}-1} \\
&\leq \frac{1}{(r+n_i-1)\cdot\ldots\cdot r} \prod_{k\in W_i} r(r+1)^{y_{ik}-1}(p_{ik}^E)^{min(y_{ik},1)} \\
&= \frac{r^{d_i}(r+1)^{n_i-d_i}}{(r+n_i-1)\cdot\ldots\cdot r} \prod_{k\in W_i} (p_{ik}^E)^{min(y_{ik},1)} \\
&\leq \frac{r}{r+1} \prod_{k\in W_i} (p_{ik}^E)^{min(y_{ik},1)},
\end{aligned}
\tag{30}
$$

up to a multiple constant.

Finally, for extreme group i ($d_i = 1$):

$$
\begin{aligned}
p(\boldsymbol{y_i}|r,\boldsymbol{\beta}) &\propto \frac{\Gamma(r)}{\Gamma(n_i+r)} \cdot \frac{\Gamma(y_{ij}+rp_{ij}^E)}{\Gamma(rp_{ij}^E)} = \frac{\Gamma(r)}{\Gamma(n_i+r)} \cdot \frac{\Gamma(n_i+rp_{ij}^E)}{\Gamma(rp_{ij}^E)} \\
&\leq \frac{rp_{ik}^E(rp_{ik}^E+1)^{n_i-1}}{(r+n_i-1)\cdot\ldots\cdot(r+1)r} \\
&\leq \frac{r(r+1)^{n_i-1}(p_{ik}^E)^{min(n_i,1)}}{(r+n_i-1)\cdot\ldots\cdot(r+1)r} \\
&\leq (p_{ik}^E)^{min(n_i,1)}
\end{aligned}
\tag{31}
$$

From all the three upper bounds (29), (30) and (31) for the Dirichlet-multinomial probability mass function with non-integer counts, the r functions are the same as in the condition of integer counts. The only difference for the β functions is that when the count $0 < y_{ik} < 1$, p_{ik}^E is replaced by $(p_{ik}^E)^{y_{ik}}$ in the upper bounds and the posterior propriety still holds when $\boldsymbol{X}_y$ is of full rank. Thus,

the sufficient condition for posterior properiety does not change for data with non-integer counts.

Bibliography

[1] Robert Ashmead and Eric Slud. Small area model diagnostics and validation with applications to the Voting Rights Act section 203. *JSM Proceedings, Survey Research and Methodology Section. Alexandria, VA: American Statistical Association*, 2017.

[2] The United States Census Bureau. Frequently occurring surnames from the 2010 Census. `https://www.census.gov/topics/population/genealogy/data/2010_surnames.html`. Accessed: 2018-07-04.

[3] The United States Census Bureau. Public use microdata areas (PUMAs). `https://www.census.gov/geo/reference/puma.html`. Accessed: 2018-07-01.

[4] The United States Census Bureau. American Community Survey 2016 ACS 1-year PUMS files readme. `https://www2.census.gov/programs-surveys/acs/tech_docs/pums/ACS2016_PUMS_README.pdf`, 2017. Accessed: 2018-07-01.

[5] The United States Census Bureau. Quickfacts United States. `https://www.census.gov/quickfacts/fact/table/US/PST045216`, 2017. Accessed: 2018-07-01.

[6] The United States Census Bureau. About the American Community Survey. `https://www.census.gov/programs-surveys/acs/about.html`, 2018. Accessed: 2018-06-30.

[7] The United States Census Bureau. Public use microdata sample (PUMS) documentation. `https://www.census.gov/programs-surveys/acs/technical-documentation/pums.html`, 2018. Accessed: 2018-07-01.

[8] The United States Census Bureau. PUMS data. `https://www.census.gov/programs-surveys/acs/data/pums.html`, 2018. Accessed: 2018-07-01.

[9] Bradley P Carlin and Thomas A Louis. *Bayes and empirical Bayes methods for data analysis*. Chapman and Hall/CRC, 2010.

[10] Ray Chambers, Nicola Salvati, and Nikos Tzavidis. Semiparametric small area estimation for binary outcomes with application to unemployment estimation for local authorities in the UK. *Journal of the Royal Statistical*

Society: Series A (Statistics in Society), 179(2):453–479, 2016.

[11] Hau-wen Chang, Dongwon Lee, Mohammed Eltaher, and Jeongkyu Lee. @ Phillies tweeting from Philly? Predicting Twitter user locations with spatial word usage. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '12, pages 111–118. IEEE Computer Society, 2012.

[12] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: a content-based approach to geo-locating Twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 759–768. ACM, 2010.

[13] Cindy L Christiansen and Carl N Morris. Hierarchical Poisson regression modeling. *Journal of the American Statistical Association*, 92(438):618–632, 1997.

[14] Cindy L Christiansen and Carl N Morris. Improving the statistical approach to health care provider profiling. *Annals of Internal Medicine*, 127(8_Part_2):764–768, 1997.

[15] Chris Elbers, Jean O Lanjouw, and Peter Lanjouw. Micro-level estimation of poverty and inequality. *Econometrica*, 71(1):355–364, 2003.

[16] Yair Ghitza and Andrew Gelman. Deep interactions with mrp: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Science*, 57(3):762–776, 2013.

[17] Malay Ghosh and JNK Rao. Small area estimation: an appraisal. *Statistical Science*, 9(1):55–76, 1994.

[18] Neung Soo Ha. *Hierarchical Bayesian estimation of small area means using complex survey data*. University of Maryland - College Park, 2013.

[19] David A Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338, 1977.

[20] Ryan Janicki and Donald Malec. A small sample evaluation of design-adjusted likelihoods using bernoulli outcomes. *Statistics*, page 05, 2014.

[21] Jiming Jiang and Partha Lahiri. Estimation of finite population domain means: A model-assisted empirical best prediction approach. *Journal of the American Statistical Association*, 101(473):301–311, 2006.

[22] Patrick M Joyce, Donald Malec, Roderick JA Little, Aaron Gilary, Alfredo Navarro, and Mark E Asiala. Statistical modeling methodology for

the Voting Rights Act section 203 language assistance determinations. *Journal of the American Statistical Association*, 109(505):36–47, 2014.

[23] Robert E Kass and Duane Steffey. Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association*, 84(407):717–726, 1989.

[24] Joseph Kelly. *Advances in the normal-normal hierarchical model*. Harvard University, 2014.

[25] Joseph Kelly, Hyungsuk Tak, and Carl N Morris. Rgbp: An r package for Gaussian, Poisson, and binomial hierarchical modeling. *Journal of Statistical Software*, 78(5):1–33, 2014.

[26] Sheila Kinsella, Vanessa Murdock, and Neil O'Hare. I'm eating a sandwich in Glasgow: modeling locations with tweets. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, SMUC '11, pages 61–68. ACM, 2011.

[27] Rui Li, Shengjie Wang, Hongbo Deng, Rui Wang, and Kevin Chen-Chuan Chang. Towards social user profiling: unified and discriminative influence model for inferring home locations. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 1023–1031. ACM, 2012.

[28] Kung-Yee Liang and Scott L Zeger. Inference based on estimating functions in the presence of nuisance parameters. *Statistical Science*, 10(2):158–173, 1995.

[29] Benmei Liu, Partha Lahiri, and Graham Kalton. Hierarchical Bayes modeling of survey-weighted small area proportions. In *Proceedings of the American Statistical Association, Survey Research Section*, pages 3181–3186, 2007.

[30] Esther López-Vizcaíno, María José Lombardía, and Domingo Morales. Small area estimation of labour force indicators under a multinomial model with correlated time and area effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(3):535–565, 2015.

[31] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Home location identification of Twitter users. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):47, 2014.

[32] Donald Malec, J Sedransk, Christopher L Moriarity, and Felicia B LeClere. Small area inference for binary variables in the National

Health Interview Survey. *Journal of the American Statistical Association*, 92(439):815–826, 1997.

[33] Momin M Malik, Hemank Lamba, Constantine Nakos, and Jurgen Pfeffer. Population bias in geotagged tweets. *People*, 1(3,759.710):3–759, 2015.

[34] Microsoft. Bing Maps. `https://www.microsoft.com/en-us/maps/choose-your-bing-maps-api`. Accessed: 2018-07-01.

[35] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and Niels J Rosenquist. Understanding the demographics of Twitter users. *ICWSM*, 11:5, 2011.

[36] Isabel Molina and JNK Rao. Small area estimation of poverty indicators. *Canadian Journal of Statistics*, 38(3):369–385, 2010.

[37] Carl N Morris. Approximating posterior distributions and posterior moments. *Bayesian Statistics*, 3:327–344, 1988.

[38] Carl N Morris and Ruoxi Tang. Estimating random effects via adjustment for density maximization. *Statistical Science*, 26(2):271–287, 2011.

[39] Monica Pratesi. *Analysis of poverty data by small area estimation*. John Wiley & Sons, 2016.

[40] Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. An analysis of the user occupational class through Twitter content. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, volume 1, pages 1754–1764. Association for Computational Linguistics, 2015.

[41] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in Twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents*, SMUC '10, pages 37–44. ACM, 2010.

[42] JNK Rao. *Small-Area Estimation*. Wiley Online Library, 2015.

[43] Kyoung Min Ryoo and Sue Moon. Inferring Twitter user locations with 10 km accuracy. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 643–648. ACM, 2014.

[44] Timo Schmid, Fabian Bruckschen, Nicola Salvati, and Till Zbiranski. Constructing sociodemographic indicators for National Statistical Institutes by using mobile phone data: estimating literacy rates in Senegal. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*,

180(4):1163–1190, 2017.

[45] Social Security. Popular names by birth year. https://www.ssa.gov/oact/babynames/. Accessed: 2018-07-01.

[46] Luke Sloan, Jeffrey Morgan, Pete Burnap, and Matthew Williams. Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. *PloS one*, 10(3):e0115545, 2015.

[47] Eric Slud and Robert Ashmead. Hybrid BRR and parametric-bootstrap variance estimates for small domains in large surveys. *JSM Proceedings, Survey Research and Methodology Section. Alexandria, VA: American Statistical Association*, 2017.

[48] Statista. Distribution of Twitter users in the United States as of January 2017, by gender. https://www.statista.com/statistics/678794/united-states-twitter-gender-distribution/. Accessed: 2018-07-01.

[49] Statista. Number of mobile phone users worldwide from 2015 to 2020 (in billions). https://www.statista.com/statistics/274774/forecast-of-mobile-phone-users-worldwide/, 2018. Accessed: 2018-07-01.

[50] Statista. Number of social media users worldwide from 2010 to 2021 (in billions). https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/, 2018. Accessed: 2018-07-01.

[51] Hyungsuk Tak and Carl N Morris. Data-dependent posterior propriety of a Bayesian beta-binomial-logit model. *Bayesian Analysis*, 12(2):533–555, 2017.

[52] Stan Development Team. RStan: the R interface to Stan. *R package version 2.14. 1*, 2016.

[53] IPUMS USA. IPUMS-USA GIS boundary files. https://usa.ipums.org/usa/volii/boundaries.shtml. Accessed: 2018-07-01.

[54] IPUMS USA. Replicate weights in the American Community Survey / Puerto Rican Community Survey. https://usa.ipums.org/usa/repwt.shtml. Accessed: 2018-07-01.