BAYESIAN UNIT-LEVEL MODELING OF NON-GAUSSIAN SURVEY DATA UNDER INFORMATIVE SAMPLING WITH APPLICATION TO SMALL AREA ESTIMATION

A Dissertation presented to
the Faculty of the Graduate School
at the University of Missouri

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

by

PAUL A. PARKER

Scott H. Holan, Dissertation Supervisor

July 2021

The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled:

BAYESIAN UNIT-LEVEL MODELING OF NON-GAUSSIAN SURVEY DATA UNDER INFORMATIVE SAMPLING WITH APPLICATION TO SMALL AREA ESTIMATION

presented by Paul A. Parker,
a candidate for the degree of Doctor of Philosophy and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Scott H. Holan

Dr. Christopher K. Wikle

Dr. Erin M. Schliep

Dr. Claire E. Altman

ACKNOWLEDGMENTS

They say it takes a village. Although the saying under question is referring to the task of raising a child, I think it is equally applicable here. I certainly did not get here alone and I would like to thank my village.

First, thank you to my dissertation committee. Dr. Chris Wikle and Dr. Erin Schliep have taught me so much. Dr. Claire Altman has given me insight into the important questions that social scientists ask. Dr. Scott Holan, my chair, deserves a special thank you for constantly providing me with the resources and mentorship necessary to complete this work.

I would also like to thank the U.S. Census Bureau for supporting this work through their dissertation fellowship program, as well as the University of Missouri Population, Education and Health Center, for their support through the Research Data Center Doctoral Fellowship. In addition, thank you to the University of Missouri Research Data Center, where much of this work was conducted. Thank you as well to Dr. Ryan Janicki for the guidance you have provided.

To my friends and family, especially my parents and sister, thank you for the continued support, even through my absence the last few years.

Lastly, thank you to my wife, Mackenzie. You have supported me every step of the way, never questioning or complaining when my work took the time that belonged to you.

TABLE OF CONTENTS

CHAPTER

APPENDIX

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

BAYESIAN UNIT-LEVEL MODELING OF NON-GAUSSIAN SURVEY DATA UNDER INFORMATIVE SAMPLING WITH APPLICATION TO SMALL AREA ESTIMATION [1]

Paul A. Parker

Scott H. Holan, Dissertation Supervisor

ABSTRACT

Unit-level models are an alternative to the traditional area-level models used in small area estimation, characterized by the direct modeling of survey responses rather than aggregated direct estimates. These unit-level approaches offer many benefits over area-level modeling, such as potential for more precise estimates, construction of estimates at multiple spatial resolutions through a single model, and elimination of the need for benchmarking techniques, among others. Furthermore, many recent surveys collect interesting and complex data types at the unit level, such as text and functional data. Yet, unit-level models present two primary challenges that have limited their widespread use. First, when surveys have been sampled in an informative manner,

[1] Support for this research at the Missouri Research Data Center (MURDC), through the University of Missouri Population, Education and Health Center Doctoral Fellowship, as well as through the U.S. Census Bureau Dissertation Fellowship Program is gratefully acknowledged. This research was partially supported by the U.S. National Science Foundation (NSF) under NSF grant SES-1853096. This article is released to inform interested parties of ongoing research and to encourage discussion. The views expressed on statistical issues are those of the authors and not those of the NSF or U.S. Census Bureau.

it is critical to account for the design in some fashion when utilizing a model at the unit level. Second, unit-level datasets are inherently much larger than area-level ones, with responses that are typically non-Gaussian, leading to computational constraints. After providing a comprehensive review on the problem of informative sampling, this dissertation provides four computationally efficient methodologies for non-Gaussian survey data under informative sampling. This methodology relies on the Bayesian pseudo-likelihood to adjust for the survey design, as well as Bayesian hierarchical modeling to characterize various dependence structures. First, a count data model is developed and applied to small area estimation of housing vacancies. Second, modeling approaches for both binary and categorical data are developed, along with a variational Bayes procedure that may be used in extremely high-dimensional settings. This approach is applied to the problem of small area estimation of health insurance rates using the American Community Survey. Third, a nonlinear model is developed to allow for complex covariates, with application to text data contained within the American National Election Studies. Finally, a model is developed for functional covariates and applied to physical activity monitor data from the National Health and Nutrition Examination Survey.

Chapter 1

Introduction[1]

1.1 Outline of the Dissertation

This dissertation is focused on unit-level modeling of survey data under informative sampling. Particular emphasis is given to the application of these methods to the problem of small area estimation (SAE). SAE techniques can be generally categorized into area-level and unit-level approaches. The majority of the literature is built around area-level techniques, however, unit-level approaches can be advantageous for many reasons discussed herein. Unit-level approaches are also accompanied by two primary challenges: the need to account for sampling design, and the complexity and scale of data at the unit level. Unit level datasets are inherently orders of magnitude larger than data at the area level. In addition, unit-level responses are frequently non-Gaussian (binary, categorical, count, etc.), contrasting with area level responses, which are often Gaussian, or may be approximated as Gaussian after an appropriate

[1] The U.S. Census Bureau DRB approval number for this chapter is CBDRB-FY19-506.

transformation. Furthermore, unit-level data may frequently allow for complex (i.e. not scalar) covariates, such as text or functional data.

The issue of accounting for sampling design in unit-level modeling has been studied to some degree. Although this is an ongoing area of research for which further work is necessary, there is currently a multitude of available approaches to the problem. In contrast to this, there has been very little research focused around computationally efficient approaches to unit-level modeling, limiting the use of these approaches in practice. To this end, the majority of this dissertation is devoted to the development of computationally efficient unit-level models that account for the underlying sampling design.

Each chapter within this dissertation is intended to be self contained. Thus, there may be some redundancy regarding to the background information presented in each chapter. Furthermore, the notation used herein is not necessarily consistent between chapters. The remainder of this chapter serves as a literature review for the problem of informative sampling and how it relates to SAE at the unit level. A simulation study that compares a subset of the literature and an application to SAE of poverty are included as well. We draw upon this literature for the development of methodology in subsequent chapters. Chapter 2 contains methodology for count data at the unit level as well as an empirical simulation study related to SAE of housing vacancy. Chapter 3 explores the case of binary and categorical response data at the unit level under informative sampling. Both an efficient Gibbs sampler as well as a variational Bayes approximation are developed. The methodology is illustrated through an empirical simulation study, as well as an extremely high-dimensional SAE application to health insurance. Chapter 4 explores the use of unit level text covariates. An efficient nonlinear model is developed and illustrated through an empirical simulation study and SAE application using data from the American National Election Studies. In Chapter 5, methodology is developed to allow for functional covariates under informative sampling. This is illustrated through an empirical simulation study as well as an application to mortality estimation using physical activity monitor (PAM) data from the National Health and Nutrition Examination Survey. This chapter departs from previous chapters in that the analysis is focused on population and/or unit level inference rather than SAE. Finally, some concluding remarks are given in Chapter 6.

1.2 Overview

Government agencies have seen an increase in demand for data products in recent years. One trend that has accompanied this demand is the need for granular estimates of parameters of interest at small spatial scales or subdomains of the finite population. Typically, sample surveys are designed to provide reliable estimates of the parameters of interest for large domains. However, for some subpopulations, the area-specific sample size may be too small to produce estimates with adequate precision. The term *small area* is used to refer to any domain of interest, such as a geographic area or demographic cross classification, for which the domain-specific sample size is not large enough for reliable direct estimation. To improve precision, model-based methods can be used to 'borrow strength,' by relating the different areas of interest through use of linking models, and by introducing area-specific random effects and covariates. The Small Area Income and Poverty Estimates (SAIPE) program, and the Small Area Health Insurance Estimates (SAHIE) program within the U. S. Census

Bureau are two examples of government programs which produce county and sub-county level estimates for different demographic cross classifications across the entire United States using small area estimation (SAE) methods (Luery, 2011; Bauder et al., 2018). It can be difficult to generate small area estimates such as these for a number of reasons, including the fact that many geographic areas may have very limited sample sizes, if they have been sampled at all.

Models for SAE may be specified either at the area level or the unit level (see Rao and Molina, 2015, for an overview of small area estimation methodology). Area-level models treat the direct estimate (for example, the survey-weighted estimate of a mean) as the response, and typically induce some type of smoothing across areas. In this way, the areas with limited sample sizes may "borrow strength" from areas with larger samples. While area-level models are popular, they are limited, in that it is difficult to make estimates and predictions at a geographic or demographic level that is finer than the level of the aggregated direct estimate.

In contrast, unit-level models use individual survey units as the response data, rather than the direct estimates. Use of unit-level models can overcome some of the limitations of area-level models, as they constitute a bottom-up approach (i.e., they utilize the finest scale of resolution of the data). Since model inputs are at the unit-level (person-level, household-level, or establishment-level), predictions and estimates can be made at the same unit-level, or aggregated up to any desired level. Unit-level modeling also has the added benefit of ensuring logical consistency of estimates at different geographic levels and/or cross tabulations. For example, model-based county estimates are forced to aggregate to the corresponding state-level estimates, eliminating the need for ad hoc benchmarking. In addition, because the full unit-level dataset is used in the modeling, rather than the summary statistics used with area-level models, there is potential for improved precision of estimated quantities.

Although unit-level models may lead to more precise estimates, that aggregate naturally across different spatial resolutions, they also introduce new challenges. Perhaps the biggest challenge is how to account for the survey design in the model. With area-level models, the survey design is incorporated into the model through specification of a sampling distribution (typically taken to be Gaussian) and inclusion of direct variance estimates. With unit-level models, accounting for the survey design is not as straightforward. One challenge is that the sample unit response may be dependent on the probability of selection, even after conditioning on the design variables. When the response variables are correlated with the sample selection variables, the sampling scheme is said to be *informative*, and in these scenarios, in order to avoid bias, it is critical to capture the sample design in the model by including the survey weights or the design variables used to construct the survey weights.

The aim of this chapter is to present a comprehensive literature review of unit-level small area modeling strategies, with an emphasis on Bayesian approaches, and to evaluate a selection of these strategies by fitting different unit-level models on both simulated data, and on real American Community Survey (ACS) micro-data, thereby comparing model-based predictions and uncertainty estimates. In this chapter we focus mainly on model specification and methods which incorporate informative sampling designs into the small area model. Some important, related issues, that will be outside the scope of this chapter include issues related to measurement error and adjustments for nonresponse. Generally, we assume that observed survey weights have been modified to take into account nonresponse. We also avoid discussion on the relative merits of frequentist versus Bayesian methods for inference. In the simulation studies and data examples given in Sections 1.8 and 1.9, we fit three unit-level small area Bayesian models, with vague, proper priors on all unknown model parameters. Inference on the finite population parameters of interest is done using the posterior mean as a point estimate, and the posterior variance as a measure of uncertainty.

Some related work includes Hidiroglou and You (2016), who present a simulation study to compare area-level and unit-level models, both fit in a frequentist setting. They fit their models under both informative and noninformative sampling, and found that overall the unit-level models lead to better interval coverage and more precise estimates. Gelman (2007) discusses poststratification using survey data, and compares the implied weights of various models including hierarchical regression. Lumley and Scott (2017) discuss general techniques for modeling survey data with included weights. They focus on frequentist pseudo-likelihood estimation as well as hypothesis testing. Chapter 7 of Rao and Molina (2015) provides an overview of some commonly used unit-level small area models. The current chapter adds to this literature by providing a comprehensive review of unit-level small area modeling techniques, with a focus on methods which account for informative sampling designs. We mainly use Bayesian methods for inference, but note that many model-based methods are general enough to be implemented in either setting, and we highlight some scenarios where Bayesian methodology may be used.

The remainder of this chapter is organized as follows. Section 1.3 introduces the sampling framework and notation to be used throughout the chapter. We aim to keep the notation internally consistent. This may lead to differences compared to the original authors' notation styles, but should lead to easier comparison across methodologies. In Section 1.4 we cover modeling techniques that assume a noninformative survey design. The basic unit-level model is introduced, as well as extensions of this model which incorporate the design variables and survey weights. Methods which allow for an informative design are then discussed, beginning in Section 1.5. Here, we discuss analytic inference of population parameters under an informative design using pseudo-likelihood methods. Extensions of the pseudo-likelihood to hierarchical, multilevel mixed models are discussed, as well as application to small area estimation problems. In Section 1.6 we focus on models that use a sample distribution that differs from the population distribution. We conclude the review component of this chapter in Section 1.7, where we will review models that are specific to a Binomial likelihood, as many variables collected from survey data are binary in nature. In Section 1.8 we compare three selected models to a direct estimator under a simulation study designed around American Community Survey (ACS) data. Specifically, this simulation examines three Bayesian methods that span different general modeling approaches (pseudo-likelihood, nonparametric regression on the weights, and differing sample/population likelihoods) with the goal of examining the utility of each approach. The Stan code used to fit these models is available at https://github.com/paparker/Unit_Level_Models. Similarly, Section 1.9 uses the same models for a poverty estimates application similar to the Small Area Income and Poverty Estimates program (SAIPE). Finally, we provide concluding remarks in Section 1.10.

1.3 Background and notation

Consider a finite population $\mathcal{U}$ of size N, which is subset into m non-overlapping domains, $\mathcal{U}_i = \{1, \ldots, N_i\}, i = 1, \ldots, m$, where $\sum_{i=1}^{m} N_i = N$. These subgroups will typically be small areas of interest, or socio-demographic cross-classifications, such as age by race by gender within the different counties. We use y_{ij} to represent a particular response characteristic associated with unit $j \in \mathcal{U}_i$, and $\boldsymbol{x}_{ij}$ a vector of predictors for the response variables y_{ij}.

Let $\boldsymbol{Z}$ be a vector of design variables which characterize the sampling process. For example, $\boldsymbol{Z}$ may contain geographic variables used for stratifying the population, or size variables used in a probability proportional to size sampling scheme. A sample $\mathcal{S} \subset \mathcal{U} = \bigcup \mathcal{U}_i$ is selected according to a known sampling design with inclusion probabilities dependent on the design variables, $\boldsymbol{Z}$. Let $\mathcal{S}_i$ denote the sampled units in small area i, and let $\pi_{ij} = P(j \in \mathcal{S}_i \mid \boldsymbol{Z})$. The inverse probability sampling weights are denoted with $w_{ij} = 1/\pi_{ij}$. We note that as analysts, we may not have access to the functional form of $P\left(j \in \mathcal{S}_i \mid \boldsymbol{Z}\right)$, and may not even have access to the design variables $\boldsymbol{Z}$, so that the only information available to us about the survey design is through the observed values of π_{ij} or w_{ij}, for the sampled units in the population. Finally, we let $D_S = \{\{y_{ij}, \boldsymbol{x}_{ij}, w_{ij}\} : j \in \mathcal{S}_i, i = 1, \ldots, m\}$ represent the observed data. This simply consists of the responses, predictors, and sampling weights for all units included in the sample. In this context, y_{ij} is random, x_{ij} is typically considered fixed and known, and w_{ij} can either be fixed or random depending on the specific modeling assumptions.

The usual inferential goal, and the main focus of this chapter, is on estimation of the small area means, $\bar{y}_i = \sum_{j \in \mathcal{U}_i} y_{ij}/N_i$, or totals, $y_i = \sum_{j \in \mathcal{U}_i} y_{ij}$. In some situations, interest could be on estimation of descriptive population parameters, such as regression coefficients in a linear model, or on estimation of a distribution function. The best predictor, $\hat{\bar{y}}_i$ of $\bar{y}_i$, under squared error loss, given the observed data D_S, is

$$\hat{\bar{y}}_i = E\left(\bar{y}_i \mid D_S\right) = \frac{1}{N_i}\sum_{j\in\mathcal{U}_i} E\left(y_{ij} \mid D_S\right) = \frac{1}{N_i}\sum_{j\in\mathcal{S}_i} y_{ij} + \frac{1}{N_i}\sum_{j\in\mathcal{S}_i^c} E\left(y_{ij} \mid D_S\right). \qquad (1.1)$$

The first term on the right hand side of (1.1) is known from the observed sample. However, computation of the conditional expectation in the second term requires specification of a model, and potentially, depending on the model specified, auxiliary information, such as knowledege of the covariates $\boldsymbol{x}_{ij}$ or sampling weights w_{ij} for the nonsampled units. For the case where the predictors $\boldsymbol{x}_{ij}$ are categorical, the assumption of known covariates for the nonsampled units is not necessarily restrictive, if the totals, $N_{i,g}$, for each cross-classification g in each of the small areas i are known. In this case, the last term in (1.1) reduces to $N_i^{-1}\sum_g (N_{i,g} - n_{i,g}) E\left(y_{ij} \mid D_S\right)$, and only predictions for each cross-classification need to be made.

The predictor given in (1.1) is general, and the different unit-level modeling methods discussed in this chapter are essentially different methods for predicting the nonsampled units under different sets of assumptions on the finite population and the sampling scheme. An entire finite population can then be generated, consisting of the observed, sampled values, along with model-based predictions for the nonsampled individuals. The small area mean can then be estimated by simply averaging appropriately over this population. If the sampling fraction n_i/N_i in each small area is small, inference using predicted values for the entire population will be nearly the same as inference using a finite population consisting of the observed values and predicted values for the nonsampled units. In this situation, it may be more convenient to use a completely model-based approach for prediction of the small area means (Battese et al., 1988).

1.4 Unweighted analysis

1.4.1 Ignorable design

First, assume the survey design is *ignorable* or *noninformative*. Ignorable designs, such as simple random sampling with replacement, arise when the sample inclusion variable I is independent of the response variable y. In this situation, the distribution of the sampled responses will be identical to the distribution of nonsampled responses. That is, if a model $f(\cdot \mid \boldsymbol{\theta})$ is assumed to hold for all nonsampled units in the population, then it will also hold for the sampled units, since the *sample distribution* of y, $f(y \mid I = 1, \boldsymbol{\theta}) = f(y \mid \boldsymbol{\theta})$ is identical to the population distribution of y. In this case, a model can be fit to the sampled data, and the fitted model can then be used directly to predict the nonsampled units, without needing any adjustments due to the survey design.

The nested error regression model or, using the terminology of Rao and Molina (2015), the basic unit-level model, was introduced by Battese et al. (1988) for estimation of small area means using data obtained from a survey with an ignorable design. Consider the linear mixed effects model

$$y_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + v_i + e_{ij}, \tag{1.2}$$

where $i = 1, \ldots, m$ indexes the different small areas of interest, and $j \in \mathcal{S}_i$ indexes the sampled units in small area i. Here, the model errors, v_i, are i.i.d. $N\left(0, \sigma_v^2\right)$ random variables, and the sampling errors, e_{ij}, are i.i.d. $N\left(0, \sigma_e^2\right)$ random variables, independent of the model errors.

Let $\boldsymbol{V}_i$ be the covariance matrix consisting of diagonal elements $\sigma_v^2 + \sigma_e^2/n_i$, and off-diagonal elements σ_v^2. Assuming (1.2) holds for the sampled units, and the variance parameters σ_v^2 and σ_e^2 are known, the best linear unbiased predictor (BLUP) of $\bar{y}_i = \sum_{j \in \mathcal{U}_i} y_{ij}/N_i$ is

$$\hat{\bar{y}}_i = \frac{1}{N_i} \sum_{j \in \mathcal{S}_i} y_{ij} + \frac{1}{N_i} \sum_{j \in \mathcal{S}_i^c} \left(\boldsymbol{x}_{ij}^T \tilde{\boldsymbol{\beta}} + \tilde{v}_i \right), \tag{1.3}$$

where

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{i=1}^m \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i \right)^{-1} \left(\sum_{i=1}^m \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{y}_i \right),$$

$\boldsymbol{X}_i$ is the $n_i \times p$ matrix with rows $\boldsymbol{x}_{ij}^T$, and $\tilde{v}_i = (\sigma_v^2/(n_i\sigma_v^2 + \sigma_e^2)) \sum_{j \in \mathcal{S}_i} (y_{ij} - \boldsymbol{x}_{ij}^T \tilde{\boldsymbol{\beta}})$. In (1.3), as in the general expression in (1.1), the unobserved y_{ij} are replaced by model predictions. Note that evaluation of (1.3) requires knowledge of the population mean, $\bar{\boldsymbol{X}}_{ip} = \sum_{j \in \mathcal{U}_i} \boldsymbol{X}_{ij}/N_i$, of the covariates.

In practice, the variance components σ_v^2 and σ_e^2 are unknown and need to be estimated. The empirical best linear unbiased predictor (EBLUP) is obtained by substituting estimates, $\hat{\sigma}_e^2$ and $\hat{\sigma}_v^2$, (typically MLE, REML, or moment estimates) of the variance components in the above expressions (Prasad and Rao, 1990). In addition, this model could easily be fit using Bayesian hierarchical modeling rather than using the EBLUP, which would incorporate the uncertainty from the variance parameters. Molina et al. (2014) developed a Bayesian version of the nested error regression model (1.2), using noninformative priors on the variance components.

The survey weights do not enter into either the nested error regression model (1.2) or the EBLUPs of the small area means (1.3). Because of this, the EBLUP is not design-consistent, unless the sampling design is self-weighting within each small area (Rao and Molina, 2015).

1.4.2 Including design variables in the model

Suppose now that the survey design is informative, so that the way in which individuals are selected in the sample depends in an important way on the value of the response variable y_{ij}. It is well established that when the survey design is informative, that ignoring the survey design and performing unweighted analyses without adjustment can result in substantial biases (Nathan and Holt, 1980; Pfeffermann and Sverchkov, 2007).

One method to eliminate the effects of an informative design is to condition on all design variables (Gelman et al., 1995, Chap. 7). To see this, decompose the response variables as $\boldsymbol{y} = (\boldsymbol{y}_s, \boldsymbol{y}_{ns})$, where $\boldsymbol{y}_s$ are the observed responses for the sampled units in the population, and $\boldsymbol{y}_{ns}$ represents the unobserved variables corresponding to nonsampled individuals. Let $\boldsymbol{I}$ be the matrix of sample inclusion variables, so that $I_{ij} = 1$ if y_{ij} is observed and $I_{ij} = 0$ otherwise. The observed data likelihood, conditional on covariate information $\boldsymbol{X}$, and model parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, is then

$$f(\boldsymbol{y}_s, \boldsymbol{I} \mid \boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \int f(\boldsymbol{y}_s, \boldsymbol{y}_{ns}, \boldsymbol{I} \mid \boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\phi}) d\boldsymbol{y}_{ns} = \int f(\boldsymbol{I} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\phi}) f(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\theta}) d\boldsymbol{y}_{ns}.$$

If $f(\boldsymbol{I} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\phi}) = f(\boldsymbol{I} \mid \boldsymbol{X}, \boldsymbol{\phi})$, the inclusion variables $\boldsymbol{I}$ are independent of $\boldsymbol{y}$, conditional on $\boldsymbol{X}$, and the survey design can be ignored. For example, if the design variables $\boldsymbol{Z}$ are included in $\boldsymbol{X}$, the ignorability condition may hold, and inference can be based on $f(\boldsymbol{y}_s \mid \boldsymbol{X}, \boldsymbol{\theta})$.

Little (2012) advocates for a general framework using unit-level Bayesian modeling that incorporates the design variables. For example, if cluster sampling is used, one could incorporate a cluster level random effect into the model, or if a stratified design is used, one might incorporate fixed effects for the strata. The idea is that when all design variables are accounted for in the model, the conditional distribution of the response given the covariates for the sampled units is independent of the inclusion probabilities. Because the model is unit-level and Bayesian, the unsampled population can be generated via the posterior predictive distribution. Doing so provides a distribution for any finite population quantity and incorporates the uncertainty in the parameters. For example, if the population response is generated at draw k of a Markov chain Monte Carlo algorithm, $\boldsymbol{y}^{(k)}$, then one has implicitly generated a draw from the posterior distribution of the population mean for a given area i:

$$\bar{y}_i^{(k)} = \frac{\sum_{j=1}^{N} y_j^{(k)} I\left(j \in \mathcal{U}_i\right)}{\sum_{j=1}^{N} I\left(j \in \mathcal{U}_i\right)}.$$

If there are K total posterior draws, one could then estimate the mean and standard error of $\bar{y}_i$ with

$$\hat{\bar{y}}_i = \frac{1}{K} \sum_{j=1}^{K} \bar{y}_i^{(k)}$$

and

$$\widehat{SE(\hat{\bar{y}}_i)} = \sqrt{Var(\hat{\bar{y}}_i)}.$$

The problem with attempting to eliminate the effect of the design by conditioning on design variables is often more of a practical one, because neither the full set of design variables, nor the functional relationship between the design and the response variables will be fully known. Furthermore, expanding the model by including sufficient design information so as to ignore the design may make the likelihood extremely complicated or even intractable.

1.4.3 Poststratification

Little (1993) gives an overview of poststratification. To perform poststratification, the population is assumed to contain m categories, or poststratification cells, such that within each category units are independent and identically distributed. Usually these categories are cross-classifications of categorical predictor variables such as county, race, and education level. When a regression model is fit relating the response to the predictors, predictions can be generated for each unit within a cell, and thus for the entire population. Importantly, any desired aggregate estimates can easily be generated from the unit level population predictions.

Gelman and Little (1997) and Park et al. (2006) develop a framework for poststratification via hierarchical modeling. By using a hierarchical model with partial pooling, parameter estimates can be made for poststratification cells without any sampled units, and variance is reduced for cells having few sampled units. Gelman and Little (1997) and Park et al. (2006) provide an example for binary data that uses the following model

$$\begin{aligned} y_{ij}|p_{ij} &\sim Bernoulli(p_{ij}) \\ \text{logit}(p_{ij}) &= \boldsymbol{x}_{ij}'\boldsymbol{\beta} \\ \boldsymbol{\beta} &= (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) \\ \boldsymbol{\beta}_k &\overset{ind}{\sim} N_{c_k}(0, \sigma_k^2 \boldsymbol{I}_{c_k}), k = 1, \ldots, K, \end{aligned} \tag{1.4}$$

where $\boldsymbol{x}_{ij}$ is a vector of dummy variables for K categorical predictor variables with c_k classes in variable k.

Bayesian inference can be performed on this model, leading to a probability, $p_{ij} = p_i, \forall j$, that is constant within each cell $i = 1, \ldots, m$. The number of positive responses within cell i can be estimated with $N_i p_i$, and any higher level aggregate estimates can be made by aggregating the corresponding cells. In some scenarios, the number of units within each cell may not be known, in which case further modeling would be necessary.

1.5 Models with survey weight adjustments

Although many of the the models in Section 1.4 can be used to handle informative sampling, they do not rely on the survey weights. In this section, we explore techniques that rely on the weights to adjust the sample likelihood.

There have been several methods proposed in the literature which make use of the nested error regression model (1.2), but which incorporate the survey weights, either as regression variables, or as adjustments to the predicted values, so as to protect against a possible informative survey design.

1.5.1 Survey weight adjustments to the basic unit-level model

Verret et al. (2015) augmented the nested error regression model (1.2), by including functions of the inclusion probabilities, $g\left(\pi_{ij}\right)$, as predictors. Care must be taken in the choice of the function g, as knowledge of the population means $\bar{G}_i = \sum_{i \in \mathcal{U}_j} g\left(\pi_{ij}\right) / N_i$ need to be known to obtain the EBLUPs from (1.3). Some suggestions for the choice of g were $g(\pi_{ij}) = \pi_{ij}$, which gives $\bar{G}_i = 1/N_i$, and $g(\pi_{ij}) = n_i/\pi_{ij}$, which gives $\bar{G}_i = n_i \sum_{j \in \mathcal{U}_i} w_{ij}/N_i$, which may be known in practice. Verret et al. (2015) reported strong performance of the EBLUP using the augmented nested error regression model, in a probability proportional to size simulation study, in terms of bias and mean squared error, for properly chosen augmenting variable g. However, some choices of g, such as $g(\pi_{ij}) = w_{ij}$, could lead to poor performance, except under non-informative sampling. Verret et al. (2015) suggested using scatter plots of residuals from the nested error regression model against different choices of augmenting variables to choose an appropriate model. An alternative to exploring a collection of augmenting variables is to estimate the functional form of g. Zheng and Little (2003) investigated nonparametric estimation of g using penalized splines, and found that predictions of small area means using this modeling framework resulted in large gains in mean squared error over the design-based estimates in their simulation studies.

You and Rao (2002) proposed a pseudo-EBLUP of the small area means $\theta_i = \bar{\boldsymbol{X}}_i^T \boldsymbol{\beta} + \nu_i$ based on the nested error regression model (1.2), which incorporates the survey weights. In their approach, the regression parameters $\boldsymbol{\beta}$ in (1.2) are estimated by solving a system of survey-weighted estimating equations

$$\sum_{i=1}^{m} \sum_{j \in \mathcal{S}_i} w_{ij} \boldsymbol{x}_{ij} \{ y_{ij} - \boldsymbol{x}_{ij}^T \boldsymbol{\beta} - \gamma_{iw} (\bar{y}_{iw} - \bar{\boldsymbol{x}}_{iw}^T \boldsymbol{\beta}) \} = 0, \tag{1.5}$$

where $\gamma_{iw} = \sigma_\nu^2 / (\sigma_\nu^2 + \sigma_\epsilon^2 \delta_i^2)$, $\delta_i^2 = \sum_{j \in \mathcal{S}_i} w_{ij}^2$, $\bar{y}_{iw} = \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij} / \sum_{j \in \mathcal{S}_i} w_{ij}$, and $\bar{\boldsymbol{x}}_{ij} = \sum_{j \in \mathcal{S}_i} w_{ij} \boldsymbol{x}_{ij} / \sum_{j \in \mathcal{S}_i} w_{ij}$. This is an example of the pseudo-likelihood approach to incorporating survey weights, which is later discussed in more detail.

The pseudo-BLUP $\tilde{\boldsymbol{\beta}}_w = \tilde{\boldsymbol{\beta}}_w(\sigma_e^2, \sigma_v^2)$ is the solution to (1.5) when the variance components σ_e^2 and σ_v^2 are known, and the pseudo-EBLUP, $\hat{\boldsymbol{\beta}}_w = \tilde{\boldsymbol{\beta}}(\hat{\sigma}_e^2, \hat{\sigma}_v^2)$, is the solution to (1.5) using plug-in estimates $\hat{\sigma}_e^2$ and $\hat{\sigma}_v^2$ of the variance components. The pseudo-EBLUP, $\hat{\theta}_i$ of the small area mean θ_i is then

$$\hat{\theta}_{iw} = \hat{\gamma}_{iw}\bar{y}_{iw} + \left(\bar{X}_i - \hat{\gamma}_{iw}\bar{x}_{iw}\right)^T \hat{\boldsymbol{\beta}}_w.$$

Similar to Battese et al. (1988), You and Rao (2002) assumed an ignorable survey design, so that the model (1.2) holds for both the sampled and nonsampled units. However, You and Rao (2002) showed that inclusion of the survey weights in the pseudo-EBLUP results in a design-consistent estimator. In addition, when the survey weights are calibrated to the population total, so that $\sum_{j\in\mathcal{S}_i} w_{ij} = N_i$, the pseudo-EBLUP has a natural benchmarking property, without any additional adjustment, in the sense that

$$\sum_{i=1}^{m} N_i\hat{\theta}_{iw} = \hat{Y}_w + \left(\boldsymbol{X} - \hat{\boldsymbol{X}}_w\right)^T \hat{\boldsymbol{\beta}}_w,$$

where $\hat{Y}_w = \sum_{i=1}^m \sum_{j\in\mathcal{S}_i} w_{ij}y_{ij}$ and $\hat{\boldsymbol{X}}_w = \sum_{i=1}^m \sum_{j\in\mathcal{S}_i} w_{ij}\boldsymbol{x}_{ij}$. That is, the weighted sum of area-level pseudo-EBLUPs is equal to a GREG estimator of the population total.

An alternative pseudo-EBLUP, which is applicable to estimation of general small area parameters beyond the small area means, was proposed in Guadarrama et al. (2018) (see also Jiang and Lahiri, 2006). Rather than use the genuine best predictor in (1.1), which conditions on all observed data, Guadarrama et al. (2018) suggested a pseudo-best predictor, which conditions only on the survey-weighted Horvitz-Thompson estimator, $\bar{y}_{iw} = \sum_{j\in\mathcal{S}_i} w_{ij}y_{ij}/\sum_{j\in\mathcal{S}_i} w_{ij}$, of the small area means. Assuming that the nested error regression model (1.2) holds for all units in the population, there is a simple, closed-form expression for the predictions of out-of-sample variables, y_{ij}, given by

$$E\left(y_{ij} \mid \bar{y}_{iw}\right) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \gamma_{iw}\left(\bar{y}_{iw} - \bar{\boldsymbol{x}}_{iw}^T\boldsymbol{\beta}\right),$$

using the same notation as in (1.5). This idea can easily be extended for prediction of general additive parameters, $H_i = \sum_{j\in\mathcal{U}_i} h(y_{ij})/N_i$, by using the conditional expectation $E\left(h(y_{ij}) \mid \bar{y}_{iw}\right)$ in place of the out-of-sample variables.

1.5.2 Pseudo-likelihood approaches

Suppose the finite population values y_i, are independent, identically distributed realizations from a known superpopulation distribution, $f_p(y \mid \boldsymbol{\theta})$. Here, for notational convenience, we use a single subscript i to index the finite population. Standard likelihood analysis for inference on $\boldsymbol{\theta}$, using only the observed sampled values, could produce asymptotically biased estimates when the sampling design is informative (Pfeffermann et al., 1998b). Pseudo-likelihood analysis, introduced by Binder (1983) and Skinner (1989), incorporates the survey weights into the likelihood for design-consistent estimation of $\boldsymbol{\theta}$.

The pseudo-log-likelihood is defined as

$$\sum_{i\in\mathcal{S}} w_i \log f_p(y_i \mid \boldsymbol{\theta}); \tag{1.6}$$

this is simply the Horvitz-Thompson estimator of the population-level log likelihood. Inference on $\boldsymbol{\theta}$ can be based on the maximizer, $\hat{\boldsymbol{\theta}}_{PS}$ (designating the maximum of the pseudo-likelihood rather than the likelihood), of (1.6), or equivalently, by solving the system

$$\sum_{i \in \mathcal{S}} w_i \frac{\partial}{\partial \boldsymbol{\theta}} \log f_p(y_i \mid \boldsymbol{\theta}) = \mathbf{0}. \tag{1.7}$$

The system (1.7) is an example of the use of survey-weighted estimating functions for inference on a superpopulation parameter (Binder, 1983; Binder and Patak, 1994). More generally, let $\boldsymbol{\theta}_N = \boldsymbol{\theta}_N\left(\{y_i\}\right)$ be a superpopulation parameter of interest, which is a function of the finite population values $y_i, i = 1, \ldots, N$, that can be obtained as a solution to a "census" estimating equation

$$\boldsymbol{\Phi}\left(\mathbf{y}; \boldsymbol{\theta}\right) = \sum_{i=1}^{N} \boldsymbol{\phi}_i(y_i; \boldsymbol{\theta}) = \mathbf{0}, \tag{1.8}$$

where the $\boldsymbol{\phi}_i$ are known functions of the data and the parameter, with mean zero under the superpopulation model. The term "census" is used to describe the estimating function (1.8), because (1.8) can only be calculated if all finite population values are observed, or if a census of the population is conducted.

The target parameter $\boldsymbol{\theta}_N$ is defined implicitly as a solution to the census estimating equation (1.8). A point estimate, $\hat{\boldsymbol{\theta}}_N$, of $\boldsymbol{\theta}_N$ can be obtained by finding a root of a design-unbiased estimate, $\hat{\boldsymbol{\Phi}}\left(\boldsymbol{y}_s; \boldsymbol{\theta}\right)$, of $\boldsymbol{\Phi}$, such as the Horvitz-Thompson estimator

$$\hat{\boldsymbol{\Phi}}\left(\boldsymbol{y}_s, \boldsymbol{\theta}\right) = \sum_{i \in \mathcal{S}} w_i \boldsymbol{\phi}_i\left(y_i; \boldsymbol{\theta}\right) = \mathbf{0}. \tag{1.9}$$

The use of an estimating function, rather than the score function, can be advantageous, as it reduces the number of assumptions about the superpopulation that need to be made. Full distributional specification is not required, and instead only assumptions about the moment structure are needed. The choice of the specific estimating function may be motivated by a conceptual superpopulation model, or a finite population parameter of interest. Regardless of whether a superpopulation model is assumed, most finite population parameters of interest can be formulated as a solution to a census estimating equation, and well-known 'model assisted' estimators of the finite population parameters can be derived as solutions of survey-weighted estimating equations. For example, the estimating function $\phi(y_i;\theta) = (y_i - \theta)$ leads to the population total, $\sum_{i\in\mathcal{U}} y_i$, and its estimator $\sum_{i\in\mathcal{S}} w_i y_i / \sum_{i\in\mathcal{S}} w_i$. If $t_x = \sum_{i\in\mathcal{U}} x_i$ is known, the pair of estimating functions $\phi_1(y_i; x_i, \theta_1) = (y_i - x_i\theta_1)$ and $\phi_2(y_i, \theta_2) = (y_i - t_x\theta_2)$, give the population total $t_y = \sum_{i\in\mathcal{U}} y_i$ and its ratio estimator $t_x \sum_{i\in\mathcal{S}} w_i y_i / \sum_{i\in\mathcal{S}} w_i x_i$. If the covariate vector $\boldsymbol{x}_i$ contains an intercept, the pair of estimating functions $\phi_1(y_i; \theta_1, \boldsymbol{\theta}_2) = (\theta_1 - t_x\boldsymbol{\theta}_2)$ and $\boldsymbol{\phi}_2 = (y_i - \boldsymbol{x}_i^T\boldsymbol{\theta}_2)\boldsymbol{x}_i$ leads to the finite population total and its GREG estimator $\sum_{i\in\mathcal{U}} \boldsymbol{x}_i^T\boldsymbol{\theta}_2$, where $\boldsymbol{\theta}_2 = (\sum_{i\in\mathcal{S}} w_i\boldsymbol{x}_i\boldsymbol{x}_i^T)^{-1}\sum_{i\in\mathcal{S}} w_i y_i\boldsymbol{x}_i$.

An important aspect of small area modeling is the introduction of area specific random effects to link the different small areas and to "borrow strength," by relating the different areas through the linking model, and introducing auxiliary covariate information, such as administrative records. The presence of random effects and the multilevel structure of small area models means that neither the pseudo-likelihood method nor the related estimating function approach can be directly applied to small area estimation problems. However, Grilli and Pratesi (2004), Asparouhov (2006), Rabe-Hesketh and Skrondal (2006) extended the pseudo-likelihood approach to accommodate models with hierarchical structure.

Let $v_i \overset{i.i.d.}{\sim} \varphi(v)$ denote the area specific random effects with common density φ. The usual choice for φ is the mean zero normal distribution with unknown variance σ^2. Suppose now that the finite population $y_{i1}, \ldots, y_{iN_i}$ in small areas $i = 1, \ldots, m$

are i.i.d. realizations from the superpopulation $f_i(y \mid \boldsymbol{\theta}, v_i)$, and let $\mathcal{S}_i$ be the sampled units in area i. The census marginal log-likelihood is obtained by integrating out the random effects from the likelihood:

$$\begin{aligned}\log L(\boldsymbol{\theta}) &= \sum_{i=1}^{m} \log \int \prod_{j\in\mathcal{U}_i} f_i(y_{ij} \mid \boldsymbol{\theta}, v)\varphi(v)dv \\ &= \sum_{i=1}^{m} \log \int \exp\left\{\sum_{j\in\mathcal{U}_i} \log f_i(y_{ij} \mid \boldsymbol{\theta}, v)\right\}\varphi(v)dv.\end{aligned} \tag{1.10}$$

Suppose the survey weights are decomposed into two components, $w_{j|i}$, and w_i, where $w_{j|i}$ is the weight for unit j in area i, given that area i has been sampled, and w_i is the weight associated to small area i. The pseudo-log-likelihood for the multilevel model can be defined by replacing $\sum_{j\in\mathcal{U}_i} \log f_i(y_{ij} \mid \boldsymbol{\theta}, v_i)$ in (1.10) by the design-unbiased estimate, $\sum_{j\in\mathcal{S}_i} w_{j|i} \log f_i(y_{ij} \mid \boldsymbol{\theta}, v_i)$, to get

$$\log \hat{L}(\boldsymbol{\theta}) = \sum_{i=1}^{m} w_i \log \int \exp\left\{\sum_{j\in\mathcal{S}_i} w_{j|i} \log f(y_{ij} \mid \boldsymbol{\theta}, v)\right\}\varphi(v)dv. \tag{1.11}$$

Analytical expressions for the maximizer of (1.11) generally do not exist, so the maximum pseudo-likelihood estimator, $\hat{\boldsymbol{\theta}}_{ps}$, must be found by numerical maximization of (1.11). Grilli and Pratesi (2004) used the `NLMIXED` procedure within SAS, using appropriately adjusted weights in the `replicate` statement and a bootstrap for mean squared error estimation. Rabe-Hesketh and Skrondal (2006) used an adaptive quadrature routine using the `gllamm` program within Stata, and derived a sandwich estimator of the standard errors, finding good coverage in their simulation studies with this estimate. Kim et al. (2017) proposed an EM algorithm for parameter estimation. Their method involves two steps, where first the random effects are treated as fixed, and a profile likelihood maximum likelihood estimator of the random effects are computed. The second step uses the EM algorithm to estimate the remaining model parameters. Their method relies on a normal approximation to the predictive distribution of the random effects, but was found to give good results with moderate cluster sizes in numerical studies. Kim et al. (2017) also gave a method for predicting random effects, using the EM algorithm and an approximating predictive distribution that was shown to be valid for sufficiently large cluster sizes, which is needed for prediction of unobserved variables.

Rao et al. (2013) noted that both design consistency and design-model consistency of $\hat{\boldsymbol{\theta}}_{ps}$ as an estimator of the finite population parameter $\boldsymbol{\theta}_N$, or the model parameter $\boldsymbol{\theta}$, respectively, requires that both the number of areas (or clusters), m, and the number of elements within each cluster, n_i, tend to infinity, and that the relative bias of the estimators can be large when the n_i are small. Rao et al. (2013) showed that consistency can be achieved with only m tending to infinity (allowing the n_i to be small) if the joint inclusion probabilities, $\pi_{jk|i}$, are available. Their method is to use the marginal joint densities $f(y_{ij}, y_{ik} \mid \boldsymbol{\theta})$, of elements in a cluster, integrating out the random effects, in the pseudo-log likelihood, and to estimate $\boldsymbol{\theta}$ by maximizing the design-weighted pseudo log likelihood

$$l_{wC}(\boldsymbol{\theta}) = \sum_{i \in \mathcal{S}} w_i \sum_{j<k \in \mathcal{S}_i} w_{jk|i} \log f(y_{ij}, y_{ik} \mid \boldsymbol{\theta}). \tag{1.12}$$

It was shown in Yi et al. (2016) that the maximizer of (1.12), $\boldsymbol{\theta}_{wC}$, is consistent for the second-level parameters $\boldsymbol{\theta}$, with respect to the joint superpopulation model and the sampling design.

There are two important considerations when using the pseudo-likelihood in multilevel models. The first is that two sets of survey weights, w_i and $w_{j|i}$ (and in the case of the method of Rao et al. (2013), higher-order inclusion probabilities, $\pi_{jk|i}$) are required, which is not typically the case; access to only the joint survey weights w_{ij} is not sufficient to use the multilevel models, unless all clusters $i = 1, \ldots, m$ sampled with certainty. The second consideration is that use of unadjusted, second level weights $w_{j|i}$ can cause significant bias in estimates of variance components. For single level models, scaling the weights by any constant factor does not change inference, as the solution to (1.7) is clearly invariant to any scaling of the weights. However, for multilevel models, the maximum pseudo-likelihood estimator and the associated mean squared prediction error may change depending on how the weights $w_{j|i}$ are scaled. Some suggestions include using scaled weights $\tilde{w}_{j|i}$ such that: 1) $\sum_{j=1}^{n_i} \tilde{w}_{j|i} = n_i$ (Asparouhov, 2006; Grilli and Pratesi, 2004; Pfeffermann et al., 1998b), 2) constant scaling across clusters, so that $\sum_{j=1}^{n_i} \tilde{w}_{j|i} = \sum_{j=1}^{n_i} w_{j|i}$ (Asparouhov, 2006), 3) $\sum_{j=1}^{n_i} \tilde{w}_{i|j} = n_i^*$, where $n_i^* = (\sum_j w_{i|j})^2 / \sum_j w_{ij}^2$ is the effective sample size in cluster i (Potthoff et al., 1992; Pfeffermann et al., 1998b; Asparouhov, 2006), and 4) unscaled (Pfeffermann et al., 1998b; Grilli and Pratesi, 2004; Asparouhov, 2006). However, Korn and Graubard (2003) showed that any scaling method can produce seriously biased variance estimates under informative sampling schemes, even with large sample sizes in each cluster. There does not seem to be a single 'best' scaling method that can be used without consideration of the sampling scheme, or the working likelihood.

The pseudo-log-likelihoods (1.6) and (1.11) suggest pseudo-likelihoods

$$\prod_{i}^{m} \prod_{j \in \mathcal{S}_i} f(y_{ij} \mid \boldsymbol{\theta})^{w_{ij}} \tag{1.13}$$

for single level models, and

$$\prod_{i=1}^{m}\left\{\int \prod_{j\in\mathcal{S}_i} f(y_{ij} \mid \boldsymbol{x}_{ij}, \boldsymbol{\theta}, v_j)^{w_{j|i}}\phi(v_j)dv_j\right\}^{w_i} \tag{1.14}$$

for multilevel models (Asparouhov, 2006). The pseudo-likelihood (4.1) is sometimes called the composite likelihood in general statistical problems, when the weights w_{ij} (not necessarily survey weights) are known positive constants, and its use is popular in problems where the exact likelihood is intractable or computationally prohibitive (Varin et al., 2011).

The pseudo-likelihood (1.13) is not a genuine likelihood, as it does not incorporate the dependence structure in the sampled data nor the relationship between the responses and the design variables beyond inclusion of the survey weights. However, the pseudo-likelihood has been shown to be a useful tool for likelihood analysis for finite population inference in both the frequentist and Bayesian framework.

By treating the pseudo-likelihood as a genuine likelihood, and specifying a prior distribution $\pi(\boldsymbol{\theta})$ on the model parameters $\boldsymbol{\theta}$, Bayesian inference can be performed on $\boldsymbol{\theta}$. For general models, Savitsky and Toth (2016) showed for certain sampling schemes, and for a class of population distributions, that the pseudo-posterior distribution using the survey weighted pseudo-likelihood, with survey weights scaled to sum to the sample size, (1.13) converges in L^1 to the population posterior distribution. This result justifies use of (1.13) in place of the likelihood in Bayesian analysis of population parameters, conditional on the observed sampled units, even when the sample design is informative. Predictions of area-level random effects as well of predictions of nonsampled units can then be made as well.

The authors focus on parameter inference and do not give any advice for making area-level estimates. However, it is straightforward to implement a model with a Bayesian pseudo-likelihood and then apply poststratification after the fact by generating the population, and thus any desired area-level estimates using (1.1). This type of pseudo-likelihood with poststratification for SAE was demonstrated in the frequentist setting by Zhang et al. (2014).

Ribatet et al. (2012) provides a discussion on the validity of Bayesian inference using the composite likelihood (1.13) in place of the exact likelihood in Bayes' formula. An example of this method used in the sample survey context can be found in Dong et al. (2014), which used a weighted pseudo-likelihood with a multinomial distribution as a model for binned response variables. They assumed an improper Dirichlet distribution on the cell probabilities, and used the associated posterior and posterior predictive distributions for prediction of the nonsampled population units.

In a similar spirit, (Rao and Wu, 2010) use a Bayesian pseudo-empirical likelihood to create estimates for a population mean. They form the pseudo-empirical likelihood function as

$$L_{PEL}(p_1, \ldots, p_n) = \prod_{i \in \mathcal{S}} p_i^{\tilde{w_i}}$$

where the weights are scaled to sum to the sample size. They accompany this with a Dirichlet prior distribution over $(p_1, \ldots, p_n)$, and thus conjugacy yields a Dirichlet posterior distribution

$$\pi(p_1, \ldots, p_n | D_S) = c(\tilde{w_1} + \alpha_1, \ldots, \tilde{w_n} + \alpha_n) \prod_{i \in \mathcal{S}} p_i^{\tilde{w_i} + \alpha_i - 1},$$

where c represents the normalizing constant. The posterior distribution of the population mean, θ, corresponds to the posterior distribution of $\sum_{i \in \mathcal{S}} p_i y_i$. It is straightforward to use Monte Carlo techniques to sample from this posterior. The authors also note that the design weights can be replaced with calibration weights in order to include auxiliary variables.

1.5.3 Regressing on the Survey Weights

Prediction of small area quantities using (1.1) requires estimation of $E(y_{ij} \mid D_s)$ for all nonsampled units in the population. One of the main difficulties in using unit-level model-based methods is the lack of knowledge of the covariates, sampling weights, or population sizes associated with the nonsampled units and small areas, that are needed to make these model-based predictions. To overcome this difficulty, Si et al. (2015) modeled the observed poststratification cells n_i, conditional on $n = \sum_{i=1}^{m} n_i$, using the multinomial distribution

$$(n_1, \ldots, n_m) \sim \text{Multinomial}\left(n; \frac{N_1/w_1}{\sum_{i=1}^{m} N_i/w_i}, \ldots, \frac{N_m/w_m}{\sum_{i=1}^{m} N_i/w_i}\right) \tag{1.15}$$

for poststratification cells $i = 1, \ldots, m$.

This model assumes that the unique values of the sample weights determine the poststratification cells, and that the sampling weight and response are the only values known for sampled units. The authors state that, in general, this assumption is untrue, because there will be cells with low probability of selection that do not show up in the sample, but the assumption is necessary in order to proceed with the model. This model yields a posterior distribution over the cell population sizes which can be used for poststratification with their response model, which uses a nonparametric

Gaussian process regression on the survey weights,

$$\begin{aligned} y_{ij}|\mu(w_i), \sigma^2 &\sim \mathrm{N}(\mu(w_i), \sigma^2) \\ \mu(w_i)|\beta, C(w_i, w_{i'}|\boldsymbol{\theta}) &\sim \mathrm{GP}(w_i\beta, C(w_i, w_{i'}|\boldsymbol{\theta})) \\ &\pi(\sigma^2, \beta, \boldsymbol{\theta}), \end{aligned} \tag{1.16}$$

for observation j in poststratification cell i. Here, $C(\cdot, \cdot|\boldsymbol{\theta})$ represents a valid covariance function that depends on parameters $\boldsymbol{\theta}$. The authors use a squared exponential function, but other covariance functions could be used in its place. The normal distribution placed over y_{ij} could be replaced with another distribution in the case of non-Gaussian data. Specifically, the authors explore the Bernoulli response case. This model implicitly assumes that units with similar weights will tend to have similar response values, which is likely not true in general. However, in the absence of any other information about the sampled units, this may be the most practical assumption. Because Si et al. (2015) assume that only the survey weights and response values are known, this methodology cannot be used for small area estimation as presented. However, the model can be extended to include other variables such as county, which would allow for area level estimation.

Vandendijck et al. (2016) extend the work of Si et al. (2015) to be applied to small area estimation. They assume that the poststratification cells are designated by the unique weights within each area. Rather than using the raw weights, they use the weights scaled to sum to the sample size within each area. They then use a similar multinomial model to Si et al. (2015) in order to perform poststratification using the posterior distribution of the poststratification cell population sizes. Assuming a

Bernoulli response, they use the model

$$\begin{aligned} y_{ij}|\eta_{ij} &\sim \text{Bernoulli}(\eta_{ij}) \\ \text{logit}(\eta_{ij}) &= \beta_0 + \mu(\widetilde{w}_{ij}) + u_i + v_i \end{aligned} \tag{1.17}$$

for unit j in small area i, with $\widetilde{w}_{ij}$ designating the scaled weights. Independent area level random effects are denoted by u_i, whereas v_i denotes spatially dependent area level random effects, for which the authors use an intrinsic conditional autoregressive (ICAR) prior. They explore the use of a Gaussian process prior over the function $\mu(\cdot)$ as well as a penalized spline approach. For their Gaussian process prior, they assume a random walk of order one. The multinomial model

$$(n_{1i}, \ldots, n_{L_i i}) \sim \text{Multinomial}\left(n_i; \frac{N_{1i}/w_{(1)i}}{\sum_{l=1}^{L_i} N_{li}/w_{(l)i}}, \ldots, \frac{N_{L_i i}/w_{(L_i)i}}{\sum_{l=1}^{L_i} N_{li}/w_{(l)i}}\right) \tag{1.18}$$

is used for poststratification, where n_{li} and N_{li} represent the known sample size and unknown population size respectively for poststrata cell l in area i. The cells are determined by the unique weights in area i, with the value of the weight represented by $w_{(l)i}$. Although Vandendijck et al. (2016) implement their model with a Bernoulli data example, this is a type of a Generalized Additive Model, and thus other response types in the exponential family may be used as well.

1.6 Likelihood-based inference using the sample distribution

The pseudo-likelihood methods discussed in Section 1.5 require specification of a superpopulation model, which is a distribution which holds for all units in the finite population. However, validating the superpopulation model based on the observed sampled values is generally not possible, unless the sampling design is not informative, in which case, the distribution for the sampled units is the same as for the nonsampled units. Under an informative sampling design, the model for the population data does not hold for the sampling data. This can be seen by application of Bayes' Theorem. Suppose the finite population values y_{ij} are independent realizations from a population with density $f_p(\cdot \mid \boldsymbol{x}_{ij}, \boldsymbol{\theta})$, conditional on a vector of covariates $\boldsymbol{x}_{ij}$, and model parameters $\boldsymbol{\theta}$. Given knowledge of this superpopulation model, as well as the distribution of the inclusion variables, the distribution of the sampled values can be derived. Define the sample density, f_s, (Pfeffermann et al., 1998a) as the density function of y_{ij}, given that y_{ij} has been sampled, that is,

$$f_s(y_{ij} \mid \boldsymbol{x}_{ij}, \boldsymbol{\theta}) = f_p(y_{ij} \mid \boldsymbol{x}_{ij}, \boldsymbol{\theta}, I_{ij} = 1) = \frac{P(I_{ij} = 1 \mid y_{ij}, \boldsymbol{x}_{ij}, \boldsymbol{\theta}) f_p(y_{ij} \mid \boldsymbol{x}_{ij}, \boldsymbol{\theta})}{P(I_{ij} = 1 \mid \boldsymbol{x}_{ij}, \boldsymbol{\theta})}. \tag{1.19}$$

From (1.19), the sample distribution differs from the population distribution, unless $P(I_{ij} = 1 \mid y_{ij}, \boldsymbol{x}_{ij}) = P(I_{ij} = 1 \mid \boldsymbol{x}_{ij})$, which occurs in ignorable sampling designs. Note that the inclusion probabilities, π_{ij}, may differ from the probabilities $P(I_{ij} = 1 \mid \boldsymbol{x}_{ij}, y_{ij}, \boldsymbol{\theta})$ in (1.19), because the latter are not conditional on the design variables $\boldsymbol{Z}$.

Equation (1.19) can be used for likelihood-based inference if the simplifying assumption that the sampled values are independent is made. While this is not true in general, asymptotic results given in Pfeffermann et al. (1998a) justify an assumption of independence of the data for certain sampling schemes when the overall sample size is large. However, direct use of (1.19) for finite population inference requires additional model specifications for the sample inclusion variables $P(I_{ij} = 1 \mid \boldsymbol{x}_{ij}, y_{ij})$ as well as $P(I_{ij} = 1 \mid \boldsymbol{x}_{ij})$. It was shown in Pfeffermann et al. (1998a) that $P(I_{ij} = 1 \mid \boldsymbol{x}_{ij}, y_{ij}) = E_P(\pi_{ij} \mid \boldsymbol{x}_{ij}, y_{ij})$, and that $P(I_{ij} = 1 \mid \boldsymbol{x}_{ij}) = E_p(\pi_{ij} \mid \boldsymbol{x}_{ij})$, so that a superpopulation model still needs to be specified for likelihood-based inference.

Ideally, one would like to specify a model for the sampled data, and to use this model fit to the sampled data to infer the nonsampled values, without specifying a superpopulation model. Pfeffermann and Sverchkov (1999) derived an important identity linking the moments of the sample and population-based moments, which allows for likelihood inference using the observed data, without explicit specification of a population model. They showed that

$$P(I_{ij} = 1 \mid y_{ij}, \boldsymbol{x}_{ij}) = E_p(\pi_{ij} \mid y_{ij}, \boldsymbol{x}_{ij}) = 1/E_s(w_{ij} \mid y_{ij}, \boldsymbol{x}_{ij}).$$

Similarly, it was shown that

$$P(I_{ij} = 1 \mid \boldsymbol{x}_{ij}) = E_p(\pi_{ij} \mid \boldsymbol{x}_{ij}) = 1/E_s(w_{ij} \mid \boldsymbol{x}_{ij}).$$

Combining these results with an application of Bayes' Theorem, as was done to arrive at (1.19), gives the distribution for the nonsampled units in the finite population

$$f_c(y_{ij} \mid \boldsymbol{x}_{ij}) \equiv f_p(y_{ij} \mid \boldsymbol{x}_{ij}, I_{ij} = 0) = \frac{E_s(w_{ij} - 1 \mid y_{ij}, \boldsymbol{x}_{ij}) f_s(y_{ij} \mid \boldsymbol{x}_{ij})}{E_s(w_{ij} - 1 \mid \boldsymbol{x}_{ij})}, \tag{1.20}$$

where f_c represents the density function of y_{ij}, given that y_{ij} has not been sampled. This result allows one to specify only a distribution for the sampled responses and a distribution for the sampled survey weights for inference on the nonsampled units, without any hypothetical distribution for the finite population. Importantly, this allows for identification of the finite population generating distribution f_p through the sample-based likelihood. It also establishes the relationship between the moments of the sample distribution and the population distribution, allowing for prediction of nonsampled units.

In the small area estimation context, the goal is prediction of the small area means, $\bar{y}_i$, which requires estimation of $E_p(y_{ij} \mid D_S)$ in (1.1) for the nonsampled units in each area i. Suppose there is an area-specific random effect, $v_i \overset{i.i.d.}{\sim} \phi(v)$, common to all units in the population in small area i, so that the population distribution can be written $f_p(y_{ij} \mid \boldsymbol{x}_{ij}, v_i, \boldsymbol{\theta})$. Pfeffermann and Sverchkov (2007) used the result in (1.20), to show how small area means can be predicted using the observed unit level data under an informative survey design. Under the assumption that $E_c(y_{ij} \mid D_s, v_i) = E_c(y_{ij} \mid \boldsymbol{x}_{ij}, v_i)$,

$$E_p(y_{ij} \mid D_s, I_{ij} = 0) = E_c(y_{ij} \mid D_s) = E_c(E_c(y_{ij} \mid \boldsymbol{x}_{ij}, v_i) \mid D_s).$$

Combining this with (1.20) allows for prediction of the small area means after specification of a model for the sampled responses, $f_s(y_{ij} \mid \boldsymbol{x}_{ij}, v_i)$, and a model for the sampled weights, $f_s(w_{ij} \mid y_{ij}, \boldsymbol{x}_{ij}, v_i)$.

The model for the survey weights can be specified conditionally on the response variables to account for the informativeness of the survey design. Possible models for the sample weights considered in the literature include the linear model (Beaumont, 2008)

$$w_{ij} = a_0 + a_1 y_{ij} + a_2 y_{ij}^2 + \boldsymbol{x}_{ij}^T \boldsymbol{\alpha} + \epsilon_{ij},$$

and the exponential model for the mean (Pfeffermann et al., 1998a; Kim, 2002; Beaumont, 2008)

$$E_s(w_{ij} \mid \boldsymbol{x}_{ij}, y_{ij}) = k_i \exp\left(a y_{ij} + \boldsymbol{x}_{ij}^T \boldsymbol{\beta}\right). \tag{1.21}$$

Pfeffermann and Sverchkov (2007) considered the case of continuous response variables, y_{ij}, and modeled the sampled response data using the nested error regression model (1.2). The exponential model for the survey weights in (1.21) was used to model the informative survey design. Under this modeling framework, they showed that the best predictor of $\bar{Y}_i$ is approximately

$$E_p(\bar{Y}_i \mid D_s) = N_i^{-1}\left[(N_i - n_i)\hat{\theta}_i + n_i\left\{\bar{y}_i + \left(\bar{\boldsymbol{X}}_i - \bar{\boldsymbol{x}}_i\right)^T \boldsymbol{\beta}\right\} + (N_i - n_i) b \sigma_e^2\right], \tag{1.22}$$

where $\hat{\theta} = \hat{u}_i + \bar{\boldsymbol{X}}_i^T \boldsymbol{\beta}$. The term $(N_i - n_i) b \sigma_e^2$ in (1.22) is an additional term from the usual best predictor in the nested error regression model (1.2), which gives a bias correction proportional to the sampling error variance σ_e^2.

León-Novelo et al. (2019) take a fully Bayesian approach by specifying a population level model for the response, $f_p(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\theta})$, as well as a population level model for the inclusion probabilities, $f_p(\pi_{ij}|y_{ij}, \boldsymbol{x}_{ij}, \boldsymbol{\theta})$. Through a Bayes rule argument similar to (1.19), they show that the implied joint distribution for the sampled units is

$$\begin{aligned} f_s(y_{ij}, \pi_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\theta}) &= f_p(y_{ij}, \pi_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\theta}, I_{ij} = 1) \\ &= \frac{\pi_{ij} f_p(\pi_{ij}|y_{ij}, \boldsymbol{x}_{ij}, \boldsymbol{\theta})}{E_{y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\theta}}\{E(\pi_{ij}|y_{ij}, \boldsymbol{x}_{ij}, \boldsymbol{\theta})\}} \times f_p(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\theta}). \end{aligned} \tag{1.23}$$

This joint likelihood for the sample can then be used in a Bayesian model by placing a prior distribution on $\boldsymbol{\theta}$. Note that $\boldsymbol{x}_{ij}$ can be split into two vectors corresponding to $f_p(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\theta})$ and $f_p(\pi_{ij}|y_{ij}, \boldsymbol{x}_{ij}, \boldsymbol{\theta})$ if desired. Consequently, the covariates for the response model and the inclusion probability model need not be the same.

Two computational concerns arise when using the likelihood as in (1.23). The first issue is that in general, the structure will not lead to conjugate full conditional distributions. To this effect, the authors recommend using the probabilistic programming language Stan (Carpenter et al. (2017)), which implements HMC for efficient mixing. The second concern is that the integral involved in the expectation term of (1.23) needs to be solved for every sampled observation at every iteration of the sampler. If the integral is intractable, it will need to be evaluated numerically, greatly increasing the necessary computation time. They show that if the lognormal distribution is used for the population inclusion probability model, then a closed form can be found for the expectation. Specifically, let $f_p(\pi_{ij}|y_{ij}, \boldsymbol{x}_{ij}, \boldsymbol{\theta}) = f(\log \pi_{ij}|\mu = y_{ij}\kappa + t(\boldsymbol{x}_{ij}, \boldsymbol{\theta}), \sigma^2 = \sigma^2_\pi)$, where $f(\cdot|\mu, \sigma^2)$ represents a normal distribution with mean μ and variance σ^2, κ is a regression coefficient, and $t(\cdot)$ some function. Then

$$f_s(y_{ij}, \pi_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\theta}) = \frac{f(\log \pi_{ij}|\mu = y_{ij}\kappa + t(\boldsymbol{x}_{ij}, \boldsymbol{\theta}), \sigma^2 = \sigma^2_\pi)}{\exp\{t(\boldsymbol{x}_{ij}, \boldsymbol{\theta}) + \sigma^2_\pi/2\} E_{y_{ij}|\boldsymbol{x}_{ij},\boldsymbol{\theta}} \left\{\exp(y_{ij}\kappa)\right\}} \times f_p(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\theta}). \quad (1.24)$$

In other words, the moment generating function of the population response model can be used to find the analytical form of the expression, as long as the moment generating function is defined on the real line. This includes important cases such as the Gaussian, Bernoulli, and Poisson distributions, which are commonly used in the context of survey data.

1.7 Binomial likelihood special cases

The special case of binary responses is of particular interest to survey statisticians, as many surveys focus on the collection of data corresponding to characteristics of sampled individuals, with a goal of estimating the population proportion or count in a small area for a particular characteristic. In this section, some techniques for modifying a working Bernoulli or binomial likelihood using unit-level weights to account for an informative sampling design are discussed.

Suppose the responses y_{ij} are binary, and the goal is estimation of finite population proportions in each of the small areas $i = 1, \dots, m$,

$$p_i = \frac{1}{N_i} \sum_{j \in \mathcal{U}_i} y_{ij}.$$

The pseudo-likelihood methods discussed in Section 1.5 can be directly applied to construct a working likelihood of independent Bernoulli distributions for the sampled survey responses as in (1.13). Zhang et al. (2014) used these ideas to fit a survey-weighted logistic regression model, with random effects included at both the county level and the state level, using the GLIMMIX procedure within SAS, to estimate chronic obstructive pulmonary disease by age race and sex categories within United States counties. Another example can be found in Congdon and Lloyd (2010), who used a Bernoulli pseudo-likelihood to estimate diabetes prevalence within U. S. states by demographic groups. Their model formulation was similar to that used by Zhang et al. (2014), but they included an additional random effect to account for spatial correlation.

Malec et al. (1999) proposed a method which is similar in spirit to the pseudo-likelihood method, which uses the survey weights to modify the shape of the binomial likelihood function. Suppose there are D demographic groups of interest and let $\mathcal{S}_d$ be the sampled individuals belonging to demographic group $d = 1, \ldots, D$. Instead of the usual independent binomial likelihood $\prod_{id} p_{id}^{m_{id}}(1-p_{id})^{n_{id}-m_{id}}$, Malec et al. (1999) proposed a sample-adjusted likelihood

$$\prod_{id} \frac{p_{id}^{m_{id}}(1-p_{id})^{n_{id}-m_{id}}}{(p_{id}/\bar{w}_{1d} + (1-p_{id})/\bar{w}_{0d})^{n_{id}}}, \tag{1.25}$$

where

$$\bar{w}_{1d} = \sum_{(i,j)\in\mathcal{S}_d} w_{ijd} y_{ijd} / \sum_{(i,j)\in\mathcal{S}_d} y_{ijd}$$

and

$$\bar{w}_{0d} = \sum_{(i,j)\in\mathcal{S}_d} w_{ijd}(1-y_{ijd}) / \sum_{(i,j)\in\mathcal{S}_d} (1-y_{ijd}).$$

The quantities $\bar{w}_{1d}$ and $\bar{w}_{0d}$ are used to represent sampling weights for a demographic group d averaged over all individuals with and without a characteristic of interest, respectively. The justification of the denominator of (1.25) as an adjustment to the likelihood to account for informative sampling is presented in Malec et al. (1999) through use of Bayes' rule and by considering the empirical distribution of the inclusion probabilities.

An alternative approach to the pseudo-likelihood method is to attempt to construct a new, approximate likelihood with independent components, which matches the information contained in the survey sample. Let

$$\hat{p}_i = \frac{\sum_{j\in\mathcal{S}_i} w_{ij} y_{ij}}{\sum_{j\in\mathcal{S}_i} w_{ij}},$$

be the direct estimate of p_i and let $\hat{V}_i$ be the estimated variances of $\hat{p}_i$. Under a simple random sampling design, the variance of the direct estimate $\hat{p}_i$ is $V_{SRS}(\hat{p}_i) = p_i(1-p_i)/n_i$, which can be estimated by $\hat{V}_{SRS}(\hat{p}_i) = \hat{p}_i(1-\hat{p}_i)/n_i$. In complex sampling designs, elements that belong to a common cluster or area may be correlated. Because of this, the information in the sample from a complex survey is not equivalent to the information in a simple random sample of the same size. The design effect for $\hat{p}_i$ is the ratio

$$d_i = d_i(\hat{p}_i) = \frac{\hat{V}_D(\hat{p}_i)}{\hat{V}_{SRS}(\hat{p}_i)} = \frac{n_i\hat{V}_D(\hat{p}_i)}{\hat{p}_i(1-\hat{p}_i)},$$

and is a measure of the extent to which the variability under the survey design differs from the variability that would be expected under simple random sampling.

The effective sample size, n_i', is defined as the ratio of the sample size to the design effect

$$n_i' = \frac{n_i}{d_i} = \frac{p_i(1-p_i)}{V_{SRS}(\hat{p}_i)}.$$

The effective sample size is an estimate of the sample size required under a non-informative simple random sampling scheme to achieve the same precision to that observed under the complex sampling design. Typically, the effective sample size n_i' will be less than n_i for complex sample designs.

Often the design effect is not available, either due to lack of available information with which to compute it, or due to computational complexity. In such cases, design weights can be used for estimation of the effective sample size. A simple estimate of the effective sample size, which uses only the design weights was derived by Kish (1965), and is given by

$$n_i' = \frac{(\sum_{j\in\mathcal{S}_i} w_{ij})^2}{\sum_{j\in\mathcal{S}_i} w_{ij}^2}.$$

Other estimates of the design effect which use the survey weights, sample sizes, and population totals, and are appropriate for stratified sampling designs, can be found in Kish (1992).

Chen et al. (2014) and Franco and Bell (2014) used the design effect and effective sample size to define the 'effective number of cases,' $y_i^* = n_i' \hat{p}_i$. The effective number of cases, y_i^*, were then modeled using a binomial, logit-normal hierarchical structure. The sample model for the effective number of cases is then

$$y_i^* \mid p_i \sim Binomial(n_i', p_i), i = 1, \ldots, m,$$

with a linking model of

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x}_i^T \boldsymbol{\beta} + v_i,$$

where the v_i are area-specific random effects. Using the effective number of cases and the effective sample size in a binomial model is an attempt to construct a likelihood which is valid under a simple random sampling design, and will produce approximately equivalent inferences as when using the exact, but possibly unknown or computationally intractable likelihood.

Different distributional assumptions on the random effects can be made to accommodate aspects of the data or different correlation structures particular to sampled geographies. Noting that it might be expected that areas which are close to each other might share similarities, Chen et al. (2014) decomposed the random effects v_i into spatial and a non-spatial components, so that $v_i = u_i + \varepsilon_i$, where $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma_\varepsilon^2)$, and

$$u_i \mid u_j, j \in ne(i) \sim N(\bar{u}_i, \sigma_u^2/n_i). \tag{1.26}$$

Here, $ne(i)$ is the set of neighbors of area i and $\bar{u}_i$ is the mean of the neighboring spatial effects. The spatial model in (1.26) is known as the intrinsic conditional autoregressive (ICAR) model (Besag, 1974).

Franco and Bell (2014) introduced a time dependence structure into the random effect v_i for situations in which there are data from multiple time periods available, and applied their model to estimation of poverty rates using multiple years of American Community Survey data. In their formulation, the random effects have an AR(1) correlation structure, so that the model becomes

$$\begin{aligned} y^*_{i,t} \mid p_{i,t} &\sim Binomial(n'_{i,t}, p_{i,t}),\ i = 1, \ldots, m,\ t = 1, \ldots, T \\ \text{logit}(p_{i,t}) &= \boldsymbol{x}_{i,t}^T \boldsymbol{\beta}_t + \sigma_t^2 v_{i,t} \\ v_{i,t} &= \phi v_{i,t-1} + \varepsilon_{i,t} \end{aligned}$$

where $|\phi| < 1$, and the $\varepsilon_{i,t}$ are assumed to be i.i.d. $N(0, 1-\phi^2)$ random variables. The unknown parameters $\boldsymbol{\beta}_t$ and σ_t^2 are allowed to vary over time. Franco and Bell (2014) showed that the reductions in prediction uncertainty can be meaningful when the autoregressive parameter ϕ is large, but that the reduction in prediction uncertainty is more modest when $|\phi| < 0.4$. As noted by Chen et al. (2014), the inclusion of spatial or spatio-temporal random effects has the added benefit that the dependent random effects can serve as a surrogates for the variables responsible for dependency in the data.

The above methods use the survey weights either to modify the shape of an independent likelihood (Malec et al., 1997; Zheng and Little, 2003) to account for the informative design, or to estimate a design effect in an attempt to match the information contained in the survey sample to the information implied by an independent likelihood by adjusting the sample size (Chen et al., 2014; Franco and Bell, 2014). Alternatively, one could specify a working independence model for the sampled units and incorporate the survey design by using the survey weights as predictors (Zheng and Little, 2003), and to induce dependence through a latent process model.

1.8 Empirical simulation study

Unit-level models offer several potential benefits (e.g., no need for benchmarking and increased precision), however, accounting for the informative design is critical at the unit-level. There are a variety of ways to approach this; however, the utility of each approach is not apparent. We choose three methods that span different general modeling approaches (pseudo-likelihood, nonparametric regression on the weights, and differing sample/population likelihoods), in order to address this question. We choose to sample a population based on existing survey data from a complicated design, and make estimates for poverty (similar to SAIPE).

To construct a simulation study, we require a population for which the response is known for every individual, in order to compare any estimates to the truth. It is also desirable to have an informative sample. We treat the 2014 ACS sample from Minnesota as our population (around 120,000 observations and 87 counties), and further sample 10,000 observations in order to generate our estimates from the selected models. Ideally, we would mimic the survey design used by ACS, however the design is highly complex which makes replication difficult. Instead, we subsample the ACS sample with probability proportional to the reported sampling weights, $w_{ij}^{(o)}$, using the Midzuno method (Midzuno, 1951) from the `sampling` package in `R` (Tillé and Matei, 2016). This results in a new set of survey weights $w_{ij}^{(n)}$, which are inversely proportional to the original weights given in the ACS sample. Sampling in this manner results in a sample for which the selection probabilities are proportional to the original sampling weights. By comparing weighted and unweighted direct estimates, we show that sampling in this way yields an informative sample. We fit three models to the newly sampled dataset, and create county level estimates of the proportion of the original ACS sample below the poverty level.

Model 1

$$\begin{aligned} y_{ij}|\boldsymbol{\beta},\boldsymbol{\mu} &\propto \text{Bernoulli}(p_{ij})^{\widetilde{w}_{ij}} \\ \text{logit}(p_{ij}) &= \boldsymbol{x}_{ij}'\boldsymbol{\beta} + u_i \\ u_i &\overset{ind}{\sim} \text{N}(0, \sigma_u^2) \\ \boldsymbol{\beta} \sim \text{N}_p(\boldsymbol{0}_p, \boldsymbol{I}_{p\times p}\sigma_\beta^2), &\quad \sigma_u \sim \text{Cauchy}^+(0, \kappa_u), \end{aligned} \tag{1.27}$$

where the weights $\widetilde{w}_{ij}$ are scaled to sum to the total sample size, as recommended by Savitsky and Toth (2016). We incorporate a vague prior distribution by setting $\sigma_\beta^2 = 10$ and $\kappa_u = 5$. This approach is based on the Bayesian pseudo-likelihood given in Savitsky and Toth (2016). The model structure is similar to that of Zhang et al. (2014), although we use the psuedo-likelihood in a Bayesian context rather than a frequentist one. Our design matrix $\boldsymbol{X}$ includes terms for age category, race category, and sex. We use poststratification by generating the nonsampled population at every iteration of our MCMC, which we use to produce our estimates based on (1.1). The poststratification cells consist of the unique combinations of county, age category, race category, and sex, for which the population sizes are known to us.

Model 2

$$
\begin{aligned}
y_{ij}|\beta_0, f(w_{ij}), \boldsymbol{u}, \boldsymbol{v} &\sim \text{Bernoulli}(p_{ij}) \\
\text{logit}(p_{ij}) &= \beta_0 + f(w_{ij}) + u_i + v_i \\
f(w_{ij})|\gamma, \rho &\sim \text{GP}(0, \text{Cov}(f(w_{ij}), f(w_{i'j'}))) \\
\text{Cov}(f(w_{ij}), f(w_{i'j'})) &= \gamma^2 \text{exp}\left(-\frac{(w_{ij} - w_{i'j'})^2}{2\rho^2}\right) \\
\boldsymbol{u}|\tau, \alpha &\sim \text{N}(0, \tau \boldsymbol{D}(\boldsymbol{I} - \alpha \boldsymbol{W})^{-1}) \\
v_i|\sigma_v^2 &\sim \text{N}(0, \sigma_v^2), \quad i = 1, \ldots, m \\
\beta_0 \sim \text{N}(0, \sigma_\beta^2), \quad \gamma &\sim \text{Cauchy}^+(0, \kappa_\gamma), \quad \rho \sim \text{Cauchy}^+(0, \kappa_\rho) \\
\tau \sim \text{Cauchy}^+(0, \kappa_\tau), \quad \alpha &\sim \text{Unif}(-1, 1), \quad \sigma_v \sim \text{Cauchy}^+(0, \kappa_v),
\end{aligned}
\tag{1.28}
$$

where $\boldsymbol{D}$ is a diagonal matrix containing the number of neighbors for each area $i = 1, \ldots, m$ and $\boldsymbol{W}$ is an area adjacency matrix. Again, we use a vague prior distribution by setting $\sigma_\beta^2 = 10$ and $\kappa_\gamma = \kappa_\rho = \kappa_\tau = \kappa_v = 5$. This is similar to the work of Vandendijck et al. (2016), but using the squared exponential covariance kernel as in Si et al. (2015), rather than a random walk prior on $f(\cdot)$. Additionally, we choose to use the conditional autoregressive structure (CAR) rather than ICAR structure on our random effects $\boldsymbol{u}$. Note that although Vandendijck et al. (2016) use the weights scaled to sum to county sample sizes as inputs into the nonparametric function $f(\cdot)$, we attained better results by using the unscaled weights. We use the multinomial model

$$(n_{1k}, \ldots, n_{L_k k}) \sim \text{Multinomial}\left(n_k; \frac{N_{1k}/w_{(1)k}}{\sum_{l=1}^{L_k} N_{lk}/w_{(l)k}}, \ldots, \frac{N_{L_k k}/w_{(L_k)k}}{\sum_{l=1}^{L_k} N_{lk}/w_{(l)k}}\right) \tag{1.29}$$

to model the population weight values, in order to perform poststratification. In this model, n_{lk} represents the sample size in poststrata cell l in area k, while N_{lk} represents the population size in the same cell. Poststratification cells are determined by unique weight values within each county, denoted $w_{(l)k}$. Because all units in the same cell will share the same weight, by determining the population size of each cell, the weights are implicitly determined, and thus the population may be generated using the model specified in (1.28).

Model 3

$$\begin{aligned} y_{ij} \mid p_{ij} &\sim \text{Bernoulli}(p_{ij}) \\ \text{logit}(p_{ij}) &= \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + u_i \\ \log(w_{ij}) \mid y_{ij} &\sim \boldsymbol{x}_{ij}^T\boldsymbol{\alpha} + y_{ij} * a + \epsilon_{ij} \\ u_i &\overset{i.i.d.}{\sim} \text{N}\left(0, \sigma_u^2\right) \\ \epsilon_{ij} &\overset{i.i.d.}{\sim} \text{N}\left(0, \sigma_\epsilon^2\right), \end{aligned} \tag{1.30}$$

with vague $\text{N}(0, 10)$ priors on the regression coefficients $\boldsymbol{\beta}, \boldsymbol{\alpha}$, and a, and vague $\text{Cauchy}^+(0, 5)$ priors on the variance components σ_u and σ_ϵ. This model acts as a Bayesian extension of Pfeffermann and Sverchkov (2007).

All 3 models were fit via HMC using Stan (Carpenter et al., 2017). We ran each model using two chains, each of length 2,000, and discarding the first 1,000 iterations as burn-in, thus using a total of 2,000 MCMC samples. Convergence was assessed visually via traceplots of the sample chains, with no lack of convergence detected. We repeated the simulation 50 times, with a sample size of 10,000 each time. That is, we create 50 distinct subsamples from the ACS sample, and fit the three models to each subsample. We compare the mean squared error (MSE), absolute bias, 95% credible interval coverage rate for county level estimates, and computation time in seconds for each model in Table 1.1. We also compare to a Horvitz-Thompson (HT) direct estimator as well as an unweighted mean (UW) direct estimator.

Each of the three model based estimators provides a substantial reduction in MSE compared to the direct estimator, with Model 3 being the best in this regard. Additionally, Model 1 gives a low bias, quite comparable to the direct estimate. Finally, we see that Model 1 requires substantially less computation time compared to the other model-based estimators, especially when comparing to Model 2. This suggests that if one wanted to scale the model to include more data, such as estimates at a national level, Model 1 may be easier to work with. Computation times will vary depending on the specific resources used, however the main focus here is the relative time between models. Additionally, this simulation illustrates that it is feasible to fit Bayesian unit-level models in practice under reasonable computation times.

In Figure 1.1 we show the average reduction in RMSE, for each county, that was attained by the three model based estimators when compared to the HT direct estimator, averaged over the 50 simulations. Counties that did not see a reduction are plotted in gray. There are some important differences between the model results here. Specifically, Model 1 achieves a reduction in nearly every county unlike the other two

Estimator	MSE	Abs Bias	CI Cov. Rate	Time
HT Direct	0.0044	0.0063	0.77	NA
Model 1	0.0017	0.0089	0.86	107
Model 2	0.0017	0.0256	0.89	6627
Model 3	0.0009	0.0172	0.94	407
UW Direct	0.0050	0.0322	0.41	NA

Table 1.1: Simulation results: MSE, absolute bias, 95% credible interval coverage rate, and computation time in seconds were averaged over 50 simulations in order to compare the direct estimator to three model based estimators and and unweighted direct estimate.

models, but Model 3 tends to achieve a greater reduction in RMSE in general when compared to Model 1.

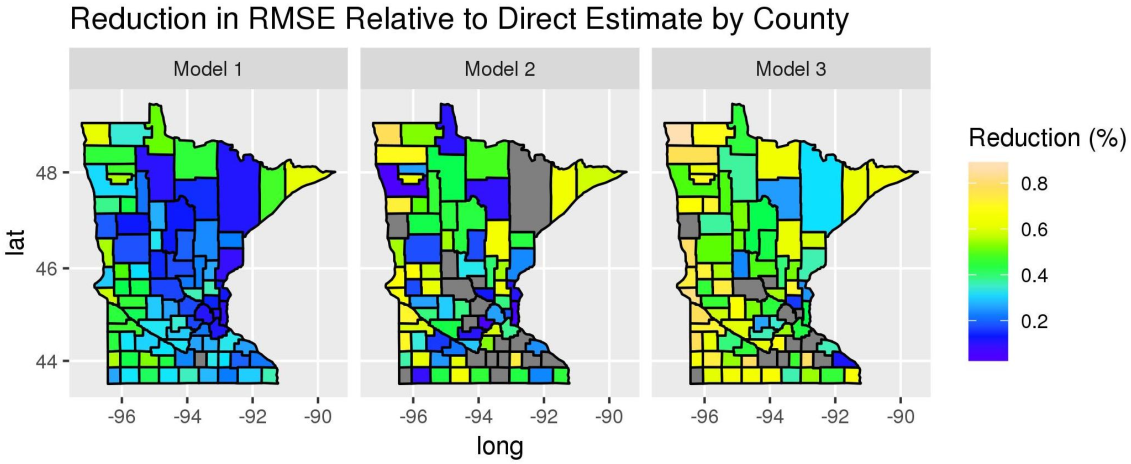

Figure 1.1: Model reduction in RMSE compared to the Direct estimates, averaged over 50 simulations. Counties that did not see a reduction are not plotted (shown in gray).

1.9 Poverty Estimate Data Analysis

The Small Area Income and Poverty Estimates program (SAIPE) is a U.S. Census Bureau program that produces estimates of median income and the number of people below the poverty threshold for states, counties, and school districts, as well as for various subgroups of the population. The SAIPE estimates are critical in order for the Department of Education to allocate Title I funds.

The current model used to generate SAIPE poverty estimates is an area-level Fay-Herriot model (Fay and Herriot, 1979) on the log scale. The response variable is the log transformed HT direct estimates from the single year ACS of the number of individuals in poverty at the county level. The model includes a number of powerful county level covariates such as the number of claimed exemptions from federal tax return data, the number of people participating in the Supplemental Nutrition Assistance Program (SNAP), and the number of Supplemental Security Income (SSI) recipients. Luery (2011) provides a comprehensive overview of the SAIPE program, including the methodology used to produce various area-level estimates and the covariates used in the model.

We use a single year of ACS data (2014 again) from Minnesota to fit the three models described in Section 1.8. The model based estimators we present are not meant to replace the current SAIPE methodology, but rather to illustrate how unit-level models can be used in an informative sampling application such as this one. The model-based predictions of the proportion of people below the poverty threshold by county under each method are presented and compared with a direct estimator.

In Figure 1.2 we show the estimate of the proportion of people below the poverty level by county for each of the model-based estimators as well as the HT direct estimator. Note that a small amount of noise has been added to the HT direct estimates as a disclosure avoidance practice. All of the estimates here seem to capture the same general spatial trend. The model based estimates resemble smoothed versions of the direct estimates, especially in the more rural areas of the state. Small sample sizes can lead to direct estimates with high variance, but the model based approaches can "share information" across areas, which leads to more precise estimates. We also compare the reduction in model based standard errors when compared to the HT direct estimate in Figure 1.3. This illustrates the precision that is gained by using a model-based estimator rather than a direct estimator in a SAE setting. Model 3 in particular appears to have the lowest standard errors in more rural areas and Model 1 seems to have lower standard errors in more populated areas. For this particular application, all three of the models we explored would be valid choices, with substantial reductions in RMSE as shown in Section 5.4.

In this case, the population cell sizes were known, however in many applications they may not be, in which case Model 2 would likely be the best option. In other cases where incorporating covariate information is desired, Model 2 is not well equipped to make estimates. This application was conducted for a single state, however if one wanted to scale the analysis, for example making estimates for every county in the United States, Model 1 appears to be the most computationally efficient. An approach similar to Model 2, albeit using a different nonlinear regression approach from the Gaussian Process regression considered here, may also be computationally efficient. Vandendijck et al. (2016) reported strong results using splines for this setup. Overall we found that each of these unit-level methods can offer precise area-level estimates, however, the properties of the particular dataset under consideration as well as the goals of the user should drive which model is selected.

Modeling poverty counts at the unit level has a number of benefits when compared to area-level models. Specifically, the current SAIPE model is on the log scale, and thus cannot naturally accommodate estimates for areas with a corresponding direct estimate of zero, whereas unit-level modeling need not be on the log scale, and thus does not suffer from this problem. Additionally, making predictions at multiple spatial resolutions is straightforward in the unit-level setting, as predictions can be generated for all units in the population and then aggregated as necessary, i.e., the so-called bottom-up approach. Under a unit-level approach, one could generate poverty estimates at both a county level and school district level under the same model. In addition to these structural benefits, Table 1.1 illustrates that unit-level models have the capacity to provide substantial reductions in MSE and variance when compared to direct estimators.

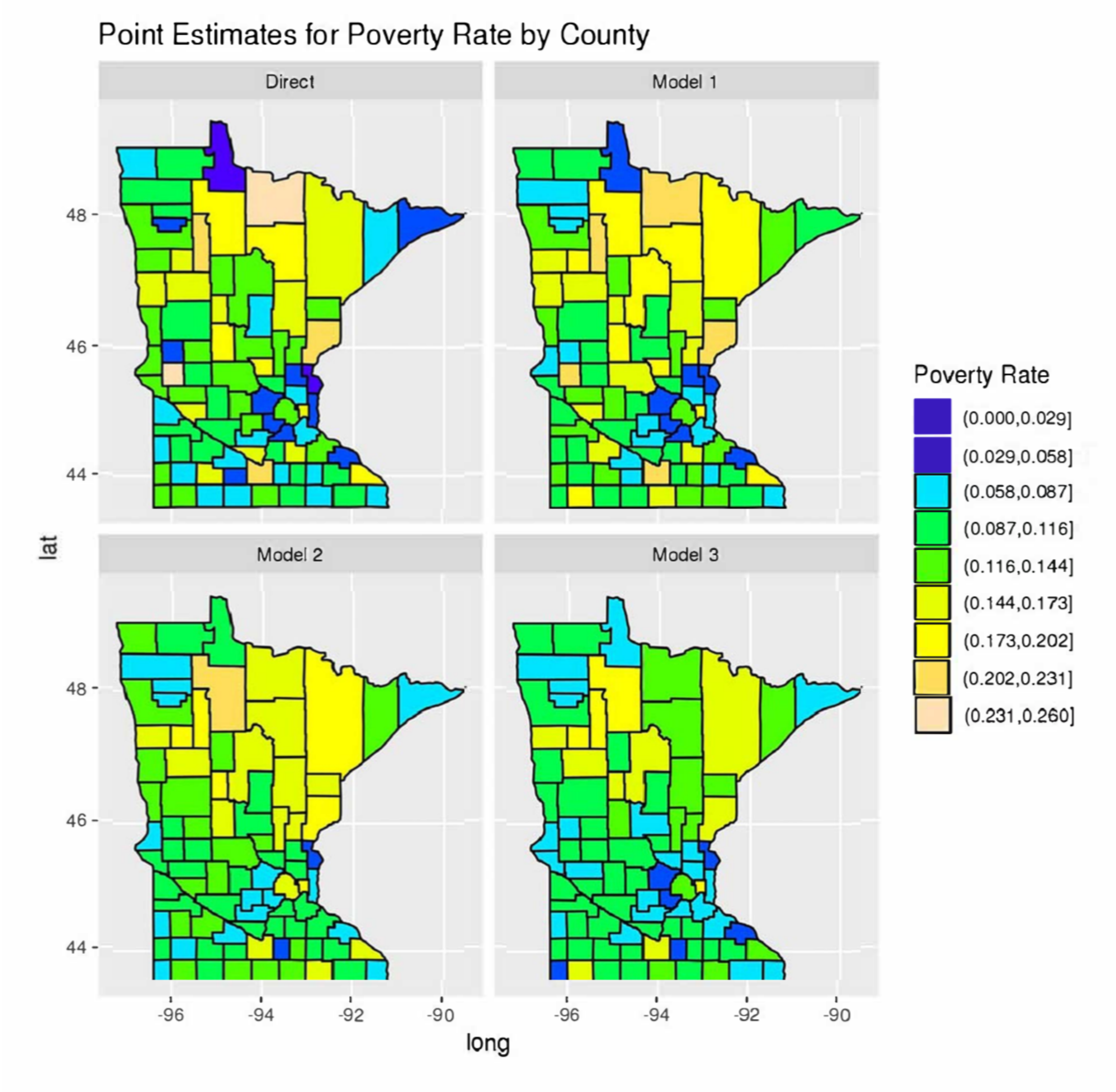

Figure 1.2: Noise infused HT direct and model based point estimates of poverty rate by county for Minnesota in 2014.

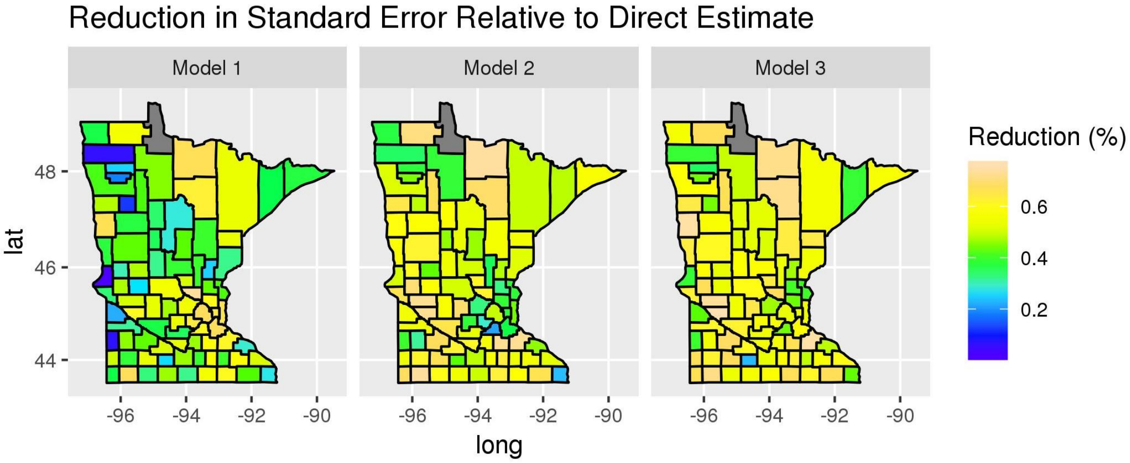

Figure 1.3: Model based reduction in standard errors for poverty rate by county. Counties that did not see a reduction are not plotted (shown in gray).

1.10 Discussion

Through a comprehensive methodological review we have demonstrated that unit-level models pose many advantages relative to area-level models. These advantages include increased precision and straightforward spatial aggregation (the so-called benchmarking problem), among others. Estimation of unit-level models requires attention to the specific sampling design. That is, the unit response may be dependent on the probability of selection, even after conditioning on the design variables. In this sense, the sampling design is said to be *informative* and care must be taken in order to avoid bias.

In the context of small area estimation, we have described several strategies for unit-level modeling under informative sampling designs and illustrated their effectiveness relative to design-based estimators (direct estimates). Specifically, our simulation study (Section 1.8) illustrated three model-based estimators that exhibited superior performance relative to the direct estimator in terms of MSE, with Model 3 performing best in this regard. Among the three models compared in this simulation, Model 1 displayed the lowest computation time relative to the other model-based estimators and, therefore, may be advantageous in higher-dimensional settings.

The models in Section 1.8 (and Section 1.9) constitute modest extensions to models currently in the literature. Specifically, Model 2 provides an extension to Vandendijck et al. (2016), whereas Model 3 can be seen as a Bayesian version of the model proposed by Pfeffermann and Sverchkov (2007). With these tools at hand, there are many opportunities for future research. For example, including administrative records into the previous model formulations constitutes one area of active research as care needs to be taken to probabilistically account for the record linkage. Methods for disclosure avoidance in unit-level models also provides another avenue for future research. In short, there are substantial opportunities for improving the models presented herein. In doing so, the aim is to provide computationally efficient estimates with improved precision. Ultimately, this will provide additional tools for official statistical agencies, survey methodologists, and subject-matter scientists.

Chapter 2

Conjugate Bayesian Unit-level Modeling of Count Data Under Informative Sampling Designs

2.1 Introduction

Statistical estimates from survey samples have traditionally been obtained via design-based estimators (Lohr, 2010). In many cases, these estimators tend to work well for quantities such as population totals or means, but can fall short as sample sizes become small. In today's "information age," there is a strong demand for more granular estimates. The Small Area Income and Poverty Estimates program (SAIPE) and the Small Area Health Insurance Estimates program (SAHIE) are two examples that rely on American Community Survey (ACS) data, where granularity is essential (Luery, 2011; Bauder et al., 2018). Both of these programs require annual estimates at the county level for the entire United States. Many counties exhibit extremely small sample sizes, or even a sample size of zero. In these cases, design-based estimation is inadequate and model-based estimation becomes necessary.

Models for survey data can be either at the area level or the unit level. Area-level models typically use design-based estimators as the response and tend to smooth the estimates in some fashion. These models often use area-level random effects to induce smoothing, and thus a common application is small area estimation (see, for example, Porter et al. (2015) and the references therein). Rao and Molina (2015) provide a recent overview of many of the current area-level models that are available. One issue with area-level models is that estimates at a finer geographic scale may not aggregate to estimates at coarser spatial resolutions, thereby producing inconsistencies.

Unit-level models include individual response values from the survey units as response variables rather than the area-level design-based estimators. The basic unit-level model was introduced by Battese et al. (1988) in order to estimate small area means. One advantage of unit-level modeling is that the response value can be predicted for all units not contained in the sample, and thus estimates for finite population quantities aggregate naturally. In addition, unit-level models have the potential to yield more precise estimates than area-level models (Hidiroglou and You, 2016). When modeling survey data at the unit level, the response is often dependent on the sample selection probabilities. This scenario is termed *informative sampling*, and it is critical to incorporate the design information into the model in order to avoid biased estimates (Pfeffermann and Sverchkov, 2007). Various approaches exist for incorporating an informative design into a model formulation. Little (2012) suggests the use of design variables in the model. For simple survey designs, this may work well, but can become infeasible for complex survey designs. Si et al. (2015) and Vandendijck et al. (2016) both use nonparametric regression techniques on the survey weights. These types of techniques do not require any knowledge of the survey design, though they can be difficult to implement in the presence of covariates. Finally, the often used pseudo-likelihood approach (Skinner, 1989; Binder, 1983) exponentially weights each unit's likelihood contribution by the corresponding survey weight. In this way, the sample data model is adjusted to better match the population distribution. See Chapter 1 for a recent review.

In addition to the issues that arise due to informative sampling, many variables found within survey data are non-Gaussian in nature, which may induce modeling difficulties. Two examples present in American Community Survey (ACS) data are a binary indicator of health insurance coverage and a count of the number of bedrooms within a household. Outside of SAE, there is a wealth of literature regarding spatial models for non-Gaussian data. For example, Diggle et al. (1998) consider the use of kriging for Poisson and Binomial count data using a spatial generalized linear mixed model. In other cases, latent Markov random fields have been used to model areal or lattice data (e.g. see Besag et al. (1991); Jin et al. (2005)). For both point and areal data, there are computationally efficient approaches available (e.g. see Banerjee et al. (2008) and Hughes and Haran (2013)). These typical spatial methods tend to model the spatial correlation at the level of the data. In contrast, unit-level SAE models are nested such that spatial dependence can be modeled between areas and units are nested within areas. The SAE setting is often aided by the use of area-level and/or unit-level random effects, which is commonly done using Bayesian hierarchical modeling with a latent Gaussian process (LGP) (Cressie and Wikle, 2011; Gelfand and Schliep, 2016). In the presence of non-Gaussian data, LGP models lead to non-conjugate full conditional distributions that can be difficult to sample from. Bradley et al. (2020) provide a solution to this problem by appealing to a class of multivariate distributions that are conjugate with members of the natural exponential family.

We introduce a modeling framework for dealing with unit-level count data under informative sampling by using Bayesian hierarchical modeling to account for complex dependence structures (e.g., in space and time), and relying on the distribution theory provided by Bradley et al. (2020) for computationally efficient sampling of the posterior distribution. To account for informative sampling, we use a Bayesian pseudo-likelihood (Savitsky and Toth, 2016). In Section 2.2 we introduce and discuss our modeling approach. Section 2.3 considers a simulation study comparing our methodology to that of two competing estimators. Finally, we provide discussion in Section 2.4.

2.2 Methodology

2.2.1 Informative Sampling

Chapter 1 reviews current approaches to unit-level modeling under informative sampling. Some of the general approaches include incorporating the design variables into the model (Little, 2012), regression on the survey weights (Si et al., 2015; Vandendijck et al., 2016), and joint modeling of the response and weights (Pfeffermann and Sverchkov, 2007; León-Novelo et al., 2019). Another general approach is to use a weighted pseudo-likelihood.

Let $\mathcal{U} = \{1, \ldots, N\}$ be an enumeration of the units in the population of interest, and let $\mathcal{S} \subset \mathcal{U}$ be the observed, sampled units, selected with probabilities $\pi_i = P(i \in \mathcal{S})$. Let y_i be a variable of interest associated with unit $i \in \mathcal{U}$. Our goal is inference on the finite population mean $\bar{y} = \sum_{i \in \mathcal{U}} y_i/n$. Suppose a model, $f(y_i \mid \theta)$, conditional on a vector of unknown parameters, θ, holds for the units y_i for $i \in \mathcal{U}$. If the survey design is informative, so that the selection probabilities, π_i, are correlated with the response variables, y_i, the model for the nonsampled units will be different from the model for the sampled units, making inference for the finite population mean challenging. Often, the reported survey weights, $w_i = 1/\pi_i$, are used to account for the survey design.

The pseudo-likelihood (PL) approach, introduced by Skinner (1989) and Binder (1983), uses the survey weights to re-weight the likelihood contribution of sampled units. The pseudo-likelihood is given by

$$\prod_{i \in \mathcal{S}} f(y_i \mid \boldsymbol{\theta})^{w_i}, \tag{2.1}$$

where y_i is the response value for unit i in the sample $\mathcal{S}$. In (2.1), the vector of model parameters is denoted by $\boldsymbol{\theta}$ and the survey weight for unit i is denoted by w_i. For frequentist estimation, the PL can be maximized via maximum likelihood techniques, whereas Savitsky and Toth (2016) use the PL in a Bayesian setting for general models. Modeling under a Bayesian pseudo-likelihood induces a pseudo-posterior distribution

$$\hat{\pi}(\boldsymbol{\theta}|\mathbf{y}, \tilde{\mathbf{w}}) \propto \left\{ \prod_{i \in \mathcal{S}} f(y_i|\boldsymbol{\theta})^{\tilde{w}_i} \right\} \pi(\boldsymbol{\theta}),$$

where $\tilde{\mathbf{w}}$ represents the weights after being scaled to sum to the sample size. This scaling is done in order to keep the asymptotic amount of information the same as the regular likelihood case, and prevent under-estimation of the standard errors, since the weights act as frequency weights (Savitsky and Toth, 2016). It was shown by Savitsky and Toth (2016), that the pseudo-posterior distribution converges to the population posterior distribution, justifying the use of the pseudo-posterior distribution for inference on the nonsampled units.

The PL approach is geared towards parameter estimates and not necessarily estimates of finite population quantities. Nevertheless, in this setting (and others), poststratification is a general technique that can be used to create finite population quantity estimates. The general idea is to use a model to predict the response value for all unsampled units in the population, effectively generating a population that can be used to extract any desired estimates. Little (1993) gives an overview of post-stratification, whereas Gelman and Little (1997) and Park et al. (2006) develop the idea of poststratification under Bayesian hierarchical models.

2.2.2 Modeling Non-Gaussian Data

Many of the variables collected from complex surveys are non-Gaussian. For example, binary indicators and count data are both very common in survey data, but cannot be modeled at the unit level under a Gaussian response framework. As such, this can lead to computational issues when dependence structures are introduced.

Bayesian hierarchical modeling is commonly used to model complex dependence structures such as those found in sample surveys. These models often consist of a data stage that models the response, a process stage, and a prior distribution over model parameters. Traditionally, a latent Gaussian process is used to model the process stage; for example see Cressie and Wikle (2011). Gelfand and Schliep (2016) review the use of Gaussian process modeling in spatial statistics, and Bradley et al. (2015) develop a general LGP framework that handles multivariate responses as well as complex spatio-temporal dependence structures.

In a Bayesian setting, when the response variable is also Gaussian, Gibbs sampling can be implemented to efficiently sample from the posterior distribution. Unfortunately, when dealing with non-Gaussian data, a Metropolis-Hastings type step may be necessary within the Markov chain Monte Carlo algorithm. Consequently, this algorithm must be tuned and can lead to poor mixing, especially in high-dimensions. Because many survey variables are inherently non-Gaussian, the Gaussian process framework is not ideal in many survey data scenarios.

Bradley et al. (2020) incorporate new distribution theory to create a set of Bayesian hierarchical models that maintain conjugacy for any response variable contained in the natural exponential family. This includes Poisson, Bernoulli, Binomial, and Gamma random variables, among others, and thus offers a very general modeling framework that maintains computational efficiency.

In this work we consider the distribution theory for Poisson responses specifically, in order to model count survey data. Bradley et al. (2020) further consider the Negative Binomial case, but state that modeling can be more challenging in this scenario. They suggest that Negative Binomial data may be alternatively modeled as Poisson, and inclusion of random effects can help to model overdispersion.

Each natural exponential family response type is shown to be conjugate with a class of distributions referred to as the conjugate multivariate (CM) distribution. For a Poisson response, the CM distribution is the multivariate log-Gamma (MLG) distribution with probability density function (PDF)

$$\det(\boldsymbol{V}^{-1})\left\{\prod_{i=1}^{n}\frac{\kappa_i^{\alpha_i}}{\Gamma(\alpha_i)}\right\}\exp\left[\boldsymbol{\alpha}'\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu})-\boldsymbol{\kappa}'\exp\left\{\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu})\right\}\right], \tag{2.2}$$

denoted by MLG$(\boldsymbol{\mu}, \mathbf{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$. The MLG distribution is easy to simulate from using the following steps:

1. Generate a vector $\mathbf{g}$ as n independent Gamma random variables with shape α_i and rate κ_i, for $i = 1, \ldots, n$
2. Let $\mathbf{g}^* = \log(\mathbf{g})$
3. Let $\mathbf{Y} = \mathbf{V}\mathbf{g}^* + \boldsymbol{\mu}$
4. Then $\mathbf{Y} \sim \text{MLG}(\boldsymbol{\mu}, \mathbf{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$.

Bayesian inference with Poisson data and MLG prior distribution also requires simulation from the conditional multivariate log-Gamma distribution (cMLG). Letting $\mathbf{Y} \sim \text{MLG}(\boldsymbol{\mu}, \mathbf{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$, Bradley et al. (2018) show that $\mathbf{Y}$ can be partitioned into $(\mathbf{Y_1}', \mathbf{Y_2}')'$, where $\mathbf{Y_1}$ is r-dimensional and $\mathbf{Y_2}$ is $(n-r)$-dimensional. The matrix $\mathbf{V}^{-1}$ is also partitioned into $[\mathbf{H}\ \mathbf{B}]$, where $\mathbf{H}$ is an $n \times r$ matrix and $\mathbf{B}$ is an $n \times (n-r)$ matrix. Then

$$\boldsymbol{Y_1}|\boldsymbol{Y_2} = \boldsymbol{d}, \boldsymbol{\mu}^*, \boldsymbol{H}, \boldsymbol{\alpha}, \boldsymbol{\kappa} \sim \text{cMLG}(\boldsymbol{\mu}^*, \boldsymbol{H}, \boldsymbol{\alpha}, \boldsymbol{\kappa}; \Psi)$$

with density

$$M\exp\left\{\boldsymbol{\alpha}'\boldsymbol{H}\boldsymbol{Y_1} - \boldsymbol{\kappa}'\exp(\boldsymbol{H}\boldsymbol{Y_1} - \boldsymbol{\mu}^*)\right\} I\left\{(\mathbf{Y}_1', \boldsymbol{d}')' \in \mathcal{M}^n\right\}, \tag{2.3}$$

where $\boldsymbol{\mu}^* = \mathbf{V}^{-1}\boldsymbol{\mu} - \mathbf{Bd}$, and M is a normalizing constant. It is also easy to sample from the cMLG distribution when doing Bayesian analysis by using a collapsed Gibbs sampler (Liu, 1994). Bradley et al. (2020) show that this can be done by drawing $(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{Y}$, where $\mathbf{Y}$ is sampled from $\mathrm{MLG}(\boldsymbol{\mu}, \mathbf{I}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$.

2.2.3 Pseudo-likelihood Poisson Multivariate log-Gamma Model

In order to use the conjugate multivariate distribution theory of Bradley et al. (2020) in a survey setting for count data under informative sampling, we replace the Poisson likelihood with a survey weighted pseudo-likelihood. Under the unweighted setting, the likelihood contribution to the posterior is proportional to

$$\prod_{i \in \mathcal{S}} \exp\left\{Z_i Y_i - b_i \exp(Y_i)\right\} = \exp\left\{\boldsymbol{Z'}\boldsymbol{Y} - \mathbf{b}'\exp(\boldsymbol{Y})\right\},$$

with $\mathbf{Z}$ representing a vector of response variables, and $\mathbf{Y}$ representing a parameter vector, which will later be modeled using the MLG distribution. The parameter $b_i = 1$ for the Poisson case. This expression is proportional to the product of Poisson densities with natural parameters $\mathbf{Y}$, $\mathrm{Pois}(\mathbf{Z}; \mathbf{Y}, \mathbf{b})$. By exponentiating the Poisson likelihood by a vector of weights, $\mathbf{W}$, the pseudo-likelihood contribution to the posterior is then proportional to

$$\prod_{i \in \mathcal{S}} \exp\left\{W_i Z_i Y_i - W_i b_i \exp(Y_i)\right\} = \exp\left\{(\boldsymbol{W} \odot \boldsymbol{Z})'\boldsymbol{Y} - (\boldsymbol{W} \odot \mathbf{b})'\exp(\boldsymbol{Y})\right\},$$

with $\odot$ representing a Hadamard product, or element-wise multiplication. This is the same form as $\mathrm{Pois}(\mathbf{Z}^*; \mathbf{Y}, \mathbf{b}^*)$, where $\mathbf{Z}^* = \mathbf{W} \odot \mathbf{Z}$ and $\mathbf{b}^* = \mathbf{W} \odot \mathbf{b}$, and thus the

MLG class of distributions is conjugate with pseudo-likelihoods built upon the Poisson distribution. This is important, as it allows us to use Gibbs sampling with conjugate full conditional distributions in order to sample from the posterior distributions.

Furthermore, Bradley et al. (2018) show that the MLG$(\mathbf{c}, \alpha^{1/2}\mathbf{V}, \alpha\mathbf{1}, \alpha\mathbf{1})$ converges in distribution to a multivariate normal distribution with mean $\mathbf{c}$ and covariance matrix $\mathbf{V}\mathbf{V}'$ as the value of α approaches infinity. This is convenient as it allows one to effectively use a latent Gaussian process model structure, while still maintaining the computationally benefits of conjugacy offered by the conjugate multivariate distribution theory. Herein, for illustration purposes, we use this type of prior distribution to approximate a latent Gaussian process. However, if desired, one could further model the shape and scale parameters from the MLG prior distribution, which can result in a more flexible shape to the posterior distribution.

We now consider the pseudo-likelihood Poisson multivariate log-Gamma model (PL-PMLG),

$$\begin{aligned}
\mathbf{Z}|\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\xi} &\propto \prod_{\ell=1}^{L}\prod_{i\in S} \text{Pois}\left(Z_i^{(\ell)}|\lambda = Y_i^{(\ell)}\right)^{\widetilde{w}_i} \\
\log(Y_i^{(\ell)}) &= \mathbf{x_i}'^{(\ell)}\boldsymbol{\beta} + \boldsymbol{\psi_i'}\boldsymbol{\eta} + \xi_i^{(\ell)},\ i \in \mathcal{S},\ \ell = 1, \ldots, L \\
\boldsymbol{\eta}|\sigma_k &\sim \text{MLG}(\mathbf{0_r}, \alpha^{1/2}\sigma_k\mathbf{I_r}, \alpha\mathbf{1_r}, \alpha\mathbf{1_r}) \\
\boldsymbol{\xi}^{(\ell)}|\sigma_\xi &\overset{ind.}{\sim} \text{MLG}(\mathbf{0_n}, \alpha^{1/2}\sigma_\xi\mathbf{I_n}, \alpha\mathbf{1_n}, \alpha\mathbf{1_n}),\ \ell = 1, \ldots, L \\
\boldsymbol{\beta} &\sim \text{MLG}(\mathbf{0_p}, \alpha^{1/2}\sigma_\beta\mathbf{I_p}, \alpha\mathbf{1_p}, \alpha\mathbf{1_p}) \\
\frac{1}{\sigma_k} &\sim \text{Log-Gamma}^{+}(\omega, \rho) \\
\frac{1}{\sigma_\xi} &\sim \text{Log-Gamma}^{+}(\omega, \rho), \quad \sigma_\beta, \alpha, \omega, \rho > 0,
\end{aligned} \tag{2.4}$$

where $Z_i^{(\ell)}$ is the ℓth response variable for unit i in the sample. This model uses a pseudo-likelihood to account for informative sampling, and is built upon a Poisson response type in order to handle count valued survey data. In this work, the vector $\boldsymbol{\psi}_i$ corresponds to an incidence vector for which areal unit i resides in. As such, the vector $\boldsymbol{\eta}$ acts as area level random effects, which are shared across response types in order to induce multivariate dependence. We note that this model is written for multivariate responses, but we focus only on a univariate example in this work. The parameters $\xi_i^{(\ell)}$ act as unit level random effects, and can account for fine scale variation due to missing unit level covariates. Finally, $\boldsymbol{\beta}$ corresponds to fixed effects, for which covariates may or may not be shared across response types. We place log-Gamma priors truncated below at zero (denoted Log-Gamma$^+$) on the parameters $1/\sigma_k$ and $1/\sigma_\xi$. This is done to maintain conjugate full conditional distributions, although other prior distributions could be used here with minimal tuning required as these are low-dimensional parameters deep in the model hierarchy. We set $\alpha = 1000$ in order to approximate Gaussian prior distributions. We also set $\sigma_\beta, \omega, \rho = 1000$ in order to create vague prior distributions. However, if prior knowledge on these parameters exists, these values could be adjusted accordingly. The full conditional distributions used for Gibbs sampling can be found in Appendix A.1.

2.2.4 Boundary Correction

One technical issue that arises when using a conjugate multivariate hierarchical modeling framework concerns data that are observed on the boundary of their support (i.e. zero counts for Poisson data). When zero counts are observed with Poisson data, the result is a full conditional distribution with a shape parameter of zero which is not well defined. Because the conjugate multivariate framework was only recently developed, there is relatively little literature on handling these boundary issues; however Bradley et al. (2020) suggest using adjusted data, $Z_i^* = Z_i + c, \;\; i = 1, \ldots, n$, by adding a small constant. This can work in many cases, depending on the dataset and the value of c, but is effectively sampling from an approximation to the posterior distribution.

Rather than sample from an approximate distribution, we use importance sampling to sample from the true posterior distribution, similar to the work of Kim et al. (1998). In this case, the importance weights are proportional to the ratio of the adjusted pseudo-likelihood to the true pseudo-likelihood. However, with large sample sizes, the adjusted pseudo-likelihood can diverge from the true pseudo-likelihood. To this effect, we run a pilot chain (using 100 iterations) to find the average ratio of the true log-pseudo-likelihood to the adjusted log-pseudo-likelihood. We then scale the weights in the adjusted pseudo-likelihood by this average ratio. This has the effect of centering the adjusted pseudo-likelihood around the true pseudo-likelihood. The importance weights, taken at each iteration of the Gibbs sampler, are then proportional to

$$\prod_{i \in \mathcal{S}} \frac{\text{Pois}(Z_i|\cdot)^{\tilde{w}_i}}{\text{Pois}(Z_i + c|\cdot)^{\tilde{w}_i^*}},$$

where $\tilde{w}_i^*$ represents the scaled survey weight after multiplying by the average ratio mentioned above. We found that for the constant c, a value of one or two was ideal, as it minimized the extent of the divergence from the true pseudo-likelihood to the approximate one.

2.3 Empirical Simulation Study

The American Community Survey (ACS) is an ongoing survey, with approximately 3.5 million households sampled annually, that is critical for informing how federal funds should be allocated. Although the complete microdata is not available to the public, public use microdata samples (PUMS) are available. PUMS only contain geographic indicators at the public use microdata area level (PUMA), which are aggregated areas such that each contains at least a population of 100,000 people. For this survey as well as others, vacant houses can pose a challenge when conducting the survey. Bradley et al. (2020) use a simple random sample of ACS PUMS data within a single PUMA to predict housing vacancies by modeling the number of people per household as a Poisson random variable. This work illustrates the capacity of unit-level models to predict housing vacancies, however, because they used simple random sampling within a single PUMA, the methodology cannot be applied in an informative sampling context for SAE.

We construct an empirical simulation study to illustrate how the PL-PMLG can be used to create small area estimates of the number of housing vacancies. Using the state of Alabama, we treat the entire 2017 PUMS housing dataset as our population (or "truth"). This dataset contains roughly 22,500 observations across 34 different PUMAs. We further subsample this data using the Midzuno probability proportional to size method (Midzuno, 1951) within the 'sampling' R package (Tillé and Matei, 2016), which we use to create our estimates. We then compare these estimates to the truth.

In addition to comparing the PL-PMLG to a direct estimator, we also wish to compare to another model based estimator. In this scenario, many of the direct estimators are equal to zero, which makes area-level modeling prohibitively difficult. Instead, because count data are often modeled as Gaussian on the log scale, we compare to a unit level model taking this approach. Because the data contains zero counts, a small constant, δ, must be added to the data before taking the log transformation, and this transformation is undone when predictions are made. The full model hierarchy, which we call the Gaussian Approximation model (GA), is

$$\begin{aligned}\log(Z_i+\delta) &\propto \mathrm{N}(\boldsymbol{x_i'\beta}+\boldsymbol{\psi_i'\eta},\sigma_\xi^2)^{\widetilde{w}_i},\ i\in\boldsymbol{\mathcal{S}}\\ \boldsymbol{\eta} &\sim \mathrm{N}(\mathbf{0},\sigma_\eta^2\boldsymbol{I})\\ \boldsymbol{\beta} &\sim \mathrm{N}(\mathbf{0},\sigma_\beta^2\boldsymbol{I})\\ \sigma_\xi^2 &\sim \mathrm{IG}(\alpha_\xi,\kappa_\xi)\\ \sigma_\eta^2 &\sim \mathrm{IG}(\alpha_\eta,\kappa_\eta)\\ \sigma_\beta^2,\alpha_\eta,&\alpha_\xi,\kappa_\eta,\kappa_\xi>0,\end{aligned} \tag{2.5}$$

where we use the vague prior distribution $\sigma_\beta^2=1000$, and $\alpha_\eta,\alpha_\xi,\kappa_\eta,\kappa_\xi=0.1$. We again use a pseudo-likelihood approach here in order to account for informative sampling. The rest of the model consists of fairly standard Bayesian mixed effects regression. We tested the value δ fixed over the values of (0.1, 1, 5), and found that $\delta=5$ yielded substantially lower MSE and bias for this example, which is what we present here.

For this simulation, we take a sample size of 5,000 from the PUMS data with probability proportional to w_i (i.e., probability inversely proportional to the original probability of selection). We show that sampling this way induces informativeness by comparing to the unweighted version of our model. Our fixed effects consist of an intercept, and the number of bedrooms in the household, which we treat as a categorical variable. We calculate the Horvitz-Thompson estimate (direct estimate) as well as the two model-based estimates. Finally, we repeat the process 50 times in order to compare MSE and absolute bias. For the PL-PMLG, unweighted PMLG (UW-PMLG) and GA estimators, we used Gibbs sampling for 2,000 iterations, discarding the first 1,000 as burn-in. Convergence was assessed visually through traceplots of the sample chains, and no lack of convergence was detected. We also compare to a Horvitz-Thompson direct estimator, with Hajek variance estimates using the `mase` package in `R` (McConville et al., 2018).

A summary of the simulation results can be found in Table 2.1, where we compare the MSE and absolute bias of the PUMA level estimates for the total number of vacant housing units. The GA model does not provide a reduction in MSE compared to the direct estimator; however, the unweighted and PL-PMLG models do (12% and 49% respectively). Additionally, the absolute bias for the PL-PMLG is substantially lower than the GA and unweighted models. The significant reduction in MSE and bias comparing the PL-PMLG and UW-PMLG models indicates that there was an informative design, and the PL approach helps to account for this design. We also show the point estimates from a randomly chosen single run of the simulation under each estimator in Figure 2.1. All of the estimators seem to capture the same general spatial trend, however the PL-PMLG estimator seems to most closely resemble the truth. As a final comparison, we plot the standard error of the estimates averaged across the 50 simulations on the log scale in Figure 2.2. To construct this figure, we compute a standard error of the estimate under each approach, for each of the 50 simulated datasets. For the model-based estimates, this standard error is the

Table 2.1: Simulation results: MSE and Absolute Bias.

Estimator	**MSE**	**Abs. Bias**
Direct	2250	3.5
GA	2526	33.9
PL-PMLG	1151	23.5
UW-PMLG	1983	32.7

posterior predictive standard deviation. We then average these standard errors across the simulated datasets, in order to illustrate the expected uncertainty associated with each reported estimate. In some cases, the standard error of the direct estimate could not be obtained due to a point estimate of zero, in which case they have been removed from the average. As expected, the standard errors are dramatically lower for the model-based estimators than the direct estimator. In general the GA standard errors are slightly lower than the PL-PMLG, however the GA exhibits much higher MSE due to the increased bias, as evidenced by Table 2.1. Thus, the PL-PMLG appears to be a superior estimator overall.

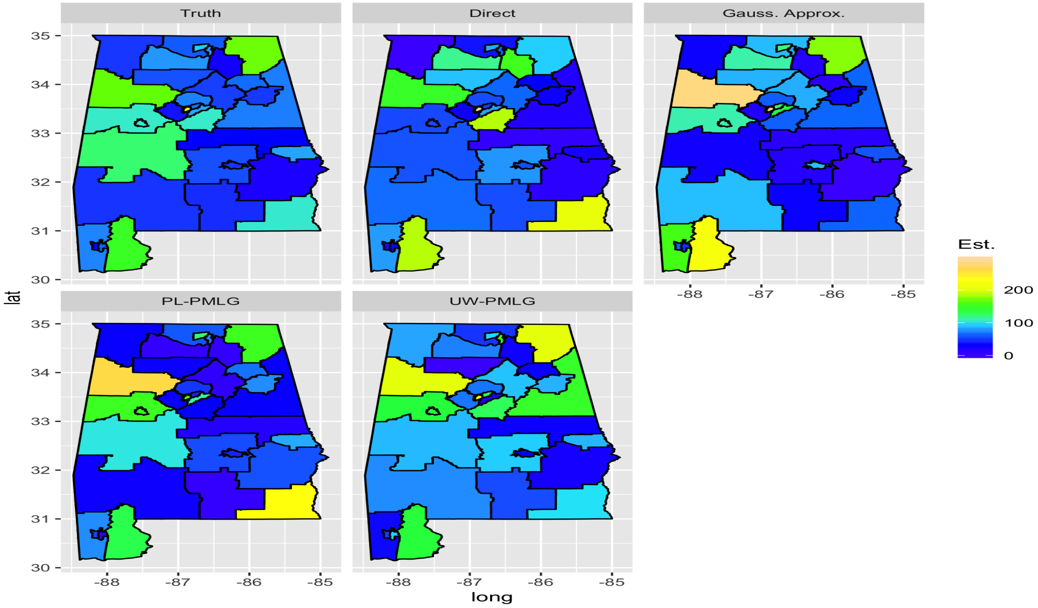

Figure 2.1: Point estimates of the number of housing vacancies by PUMA based on a single run of the simulation study.

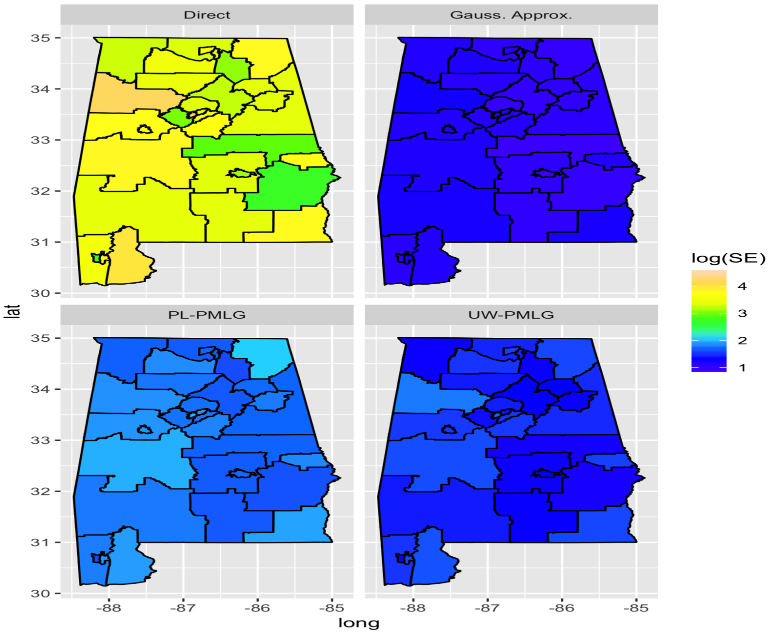

Figure 2.2: Standard error for the estimate of the number of housing vacancies by PUMA averaged over the simulation runs.

2.4 Discussion

There is a strong need for unit-level models that can handle survey data. Accounting for informative sampling design and modeling non-Gaussian data types are two of the biggest challenges in this setting. In this work, we present a new method for modeling count data while accounting for informative sampling. This method can be used for SAE as well as for more general modeling purposes. Our method relies on conjugate multivariate distribution theory, and we show that conjugacy is maintained when using a psuedo-likelihood approach to account for the survey design. We also extend the work of Bradley et al. (2020) to handle the issue of zero counts through importance sampling.

Our approach is illustrated on a simulation study built upon public-use ACS data. This is a count data example where area-level models are not feasible and Gaussian models are not appropriate. Furthermore, this is an example where direct estimators are not useful due to excessively large MSE and standard errors. Our PL-PMLG approach is able to accurately estimate population quantities based on count variables while still maintaining computational efficiency.

There still remains further work to be done in the area of non-Gaussian survey data. Other data types such as binary random variables are prevalent and should be considered (Bauder et al., 2018; Luery, 2011). The conjugate multivariate framework offered by Bradley et al. (2020) has the potential to fit these types of data, although the boundary value issue may pose a computational challenges. Our importance sampling approach works well in the Poisson case, but a more general solution may be attainable. Finally, non-Gaussian data should be explored in regards to other solutions to the informative sampling problem. The pseudo-likelihood approach may be one of the most popular approaches to informative sampling, but other methods exist and may yield additional gains in terms of precision.

Chapter 3

Computationally Efficient Bayesian Unit-Level Models for Non-Gaussian Data Under Informative Sampling[1]

3.1 Introduction

An important dichotomy in the realm of small area estimation is that of area-level versus unit-level modeling approaches. In general, area-level models use the design-based direct estimate as a response within a statistical model. These models tend to smooth the noisy direct estimates in some fashion and estimate the true latent population value. In contrast to this, unit-level models treat the individual survey respondents as observations in the statistical model. Predictions can then be made for the entire population and aggregated as necessary to produce the desired estimates.

[1] The U.S. Census Bureau DRB approval number for this paper is CBDRB-FY20-355.

As the need for more granular estimates becomes essential, area-level models may perform poorly due to underlying direct estimates with extremely small or nonexistent sample sizes. Unit-level approaches offer an attractive alternative by modeling the individual survey responses directly rather than smoothing the direct estimators. Although unit-level methodologies offer many advantages over their area-level counterparts, they also face their own set of challenges.

The primary difficulty with modeling survey data at the unit level is the consideration of informative sampling. Many surveys are sampled in an informative manner, whereby there is dependence between the probability of selection and the response of interest. When this relationship is not accounted for, increased bias may be present in the corresponding estimates (Pfeffermann and Sverchkov, 2007). The basic unit-level model, introduced by Battese et al. (1988), assumes that the sample model holds for the entire population, and thus does not account for informative sampling. Chapter 1 reviews the current methods for addressing the problem of informative sampling. Of primary interest is the pseudo-likelihood (PL) method (Skinner, 1989; Binder, 1983), which exponentially weights each unit's likelihood contribution according to the corresponding survey weight. Savitsky and Toth (2016) extend the PL approach to Bayesian settings and provide theoretical justification. Other methods to account for the survey design include modeling the design variables (Little, 2012), nonlinear regression on the survey weights (Si et al., 2015; Vandendijck et al., 2016), as well as specifying a sample model and weight model to find the implied population model (Pfeffermann and Sverchkov, 2007).

Although the problem of informative sampling has been studied in depth, there are other concerns with unit-level modeling that have received considerably less attention. In general, one major difference between area and unit-level approaches is dimensionality. Modeling survey data at the unit level can result in sample sizes that are magnitudes larger than those considered at the area level. Unit-level models are fit to individual survey responses, which can number in the millions for large-scale surveys. In contrast, area-level models are typically fit to aggregated survey statistics, such as survey-weighted means, which may number in the thousands. For example, the American Community Survey (ACS) samples 3.5 million households annually, which may reasonably fall under the realm of "big data." With these extremely large sample sizes comes computational concerns that must be addressed in order to make unit-level modeling viable. To further exacerbate the problem, many survey variables are non-Gaussian, which can lead to non-conjugate full conditional distributions when modeling dependence relationships using traditional Bayesian hierarchical models. Sampling from these posterior distributions can require Metropolis steps that are not efficient and can be cumbersome to tune.

Bradley et al. (2020) introduce a class of conjugate prior distributions that may be used to model dependence for non-Gaussian data in the natural exponential family. This covers important cases such as Binomial, Multinomial, and Poisson data. Chapter 2 extends this approach to model count data at the unit level under informative sampling, through the use of a PL. Unfortunately, sampling from the full conditional distributions can be difficult under these approaches when observations fall on the boundary of the data (i.e. zero for Poisson data, zero or one for Bernoulli data, etc.). Chapter 2 works around this by using an importance sampling scheme that works well when there are not an excessive number of boundary values (zeroes for Poisson data). However, many surveys contain a multitude of Binomial or Bernoulli random variables, which results in an abundance of boundary counts.

There are a number of data augmentation approaches that have been developed to yield conjugate full-conditional distributions for Bernoulli data. Albert and Chib (1993) use latent Gaussian variables in conjunction with a probit link function to model Bernoulli data. More recently, Polson et al. (2013) use latent Pólya-Gamma random variables to model Binomial data with a logit link function. This approach may also be used to model Negative Binomial as well as Multinomial data.

In this chapter, we develop methodology to model Binomial and Multinomial data at the unit level in a computationally efficient manner, while accounting for informative sampling. This is done through the use of Bayesian hierarchical modeling, in order to capture various sources of dependence. As previously alluded to, the weights are essential to account for the sampling design and without them we could end up with significantly biased estimates of the target tabulations. Conversely, depending on the problem, using the pseudo-likelihood can still result in a computationally intensive estimation problem. For this reason, we develop a Variational Bayes approach based on using fixed weights from the survey provided by the Official Statistical Agency. As such, we consider both a Gibbs sampling approach with fully conjugate full conditional distributions, as well as a Variational Bayes approach to model fitting.

As a motivating example, we consider the problem of estimation of the proportion of people with health insurance at the county level for different income to poverty ratio (IPR) categories. Currently, the Small Area Health Insurance Estimates (SAHIE) program within the U.S. Census Bureau produces estimates of health insurance rates using an area-level small area model fit to direct survey estimates using ACS data (Bauder et al., 2018). The model-based estimates produced by SAHIE are the only source of single year health insurance coverage estimates at the county level. While the estimates are generally more precise than the corresponding direct estimates, there are serious modeling challenges with developing area level models for health insurance coverage. First, there are boundary issues, in that many of the direct estimates at the county level are exactly equal to either 0 or 1, making use of continuous models impossible. Second, there are policy requirements to benchmark lower-level county estimates to state-level estimates, so that users have confidence in the quality of the data. Third, there are multiple within-county estimates that need to be produced, such as health insurance coverage by income level, and accounting for within-county dependencies in an area-level model can be difficult. Finally, the computational requirements of fitting the model used by SAHIE are enormous, due to the complexity of the model and the number of estimates that are produced, despite the fact that an area-level model is used.

The model proposed in this chapter eliminates many of these problems. The boundary issues are resolved by using non-Gaussian likelihoods at the unit level. There is no need to benchmark estimates, as the PL produces predictions at the unit level, which can then be aggregated up to any desired geographic level. Spatial and multivariate dependencies are handled through careful specification of the process model. Finally, computational efficiency is achieved through a Variational Bayes approximation. This work builds upon Zhang et al. (2014), who use a pseudo-likelihood for binary data in a frequentist context. In particular, we extend to the multinomial setting, which allows for categorical response data, as well as to the Bayesian pseudo-likelihood, which allows for straightforward uncertainty quantification. This chapter provides several contributions to the existing literature. Importantly, our unit-level model provides a multi-scale approach, bringing in spatial dependence at the area level, while modeling unit-level responses. Also, through the use of our multinomial specification, we are able to seamlessly combine multiple responses into one coherent modeling framework. In terms of computation, we develop a Gibbs sampling approach to model fitting, through the use of Pólya-Gamma data augmentation, building upon Polson et al. (2013). Finally, we extend the Variational Bayes approach of Durante et al. (2019), which is intended for logistic regression, to be used in the case of our pseudo-likelihood mixed model.

The remainder of this chapter is organized as follows. Section 3.2 introduces some necessary background material and then presents our proposed models as well as the methodology used to fit the models. We conduct an empirical simulation study in Section 3.3. We also provide a data analysis in Section 3.4 where we estimate the health insurance rate for each county and five different income categories for the entire continental U.S. Finally, we provide concluding remarks and discussion in Section 3.5. Although the data used herein is confidential microdata, we provide code and an example using ACS public-use microdata at `https://github.com/paparker/Unit_Level_Non-Gaussian`.

3.2 Methodology

Let $\mathcal{U} = \{1, \ldots, N\}$ be an enumeration of a finite population of interest. Suppose the finite population, $\mathcal{U}$, can be represented as the union of m non-overlapping sub-populations, or small areas, $\mathcal{U}_j = \{1, \ldots, N_j\}$, where $\sum_{j=1}^{m} N_j = N$, and $j \in \{1, \ldots, m\}$ indexes the small areas. Associated to each unit $i \in U_j$ is a characteristic of interest, Z_{ij}, and a vector, $\boldsymbol{x}_{ij}$, of auxiliary information.

A sample, $\mathcal{S} \subset \mathcal{U}$, is selected from the finite population according to a known sampling design. Let $\pi_i = P\left(i \in \mathcal{S}\right)$ be the sample inclusion probability for unit i in the finite population, and let $w_i = 1/\pi_i$ be the survey weight. A typical inferential goal is estimation of the finite population means

$$\bar{Z}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} Z_{ij} \tag{3.1}$$

from the observed survey responses. The Horvitz-Thompson estimator (Horvitz and Thompson, 1952)

$$\hat{\bar{Z}}_j = \frac{1}{N_j} \sum_{i \in \mathcal{S}_j} w_{ij} Z_{ij}, \tag{3.2}$$

where $\mathcal{S}_j = \mathcal{S} \cap \mathcal{U}_j$, is a design-unbiased and design-consistent estimator of the finite population mean, $\bar{Z}_j$. We refer to any estimate which only used the observed survey data, such as (3.2), as a *direct estimate*.

Let n be the total number of sampled units, and let n_j be the number of sampled units in $\mathcal{S}_j$. In many surveys, the overall sample size, n, and many of the area-specific sample sizes, n_j, are large. For these areas, the large-sample properties of the Horvitz-Thompson estimator guarantee that (3.2) will be a precise estimator of the finite population mean (3.1). However, it is also often the case that for many of the small areas of interest, that n_j will be too small for (3.2) to be reliable. In such situations, precision can be increased by using models for the survey data which incorporate auxiliary information to "borrow strength" by relating the different small areas and increasing the effective sample sizes.

Models for small area estimation (SAE) often include area-level random effects in order to link the small areas and incorporate spatial dependence. These random effects are typically modeled using a latent Gaussian process (LGP), and Bayesian hierarchical modeling is a common technique used to fit these models. This may be computationally efficient when considering a Gaussian response, as it leads to conjugate full conditional distributions, however when the data model (likelihood) is non-Gaussian, sampling from the posterior distribution can become difficult as it may require the use of Metropolis type steps. These sampling mechanisms require tuning that can become unwieldy especially in high dimensional situations.

Polson et al. (2013) use a data augmentation scheme to allow for conjugate sampling under logistic likelihoods. Importantly, this includes both Bernoulli and Multinomial responses, which is useful as binary and categorical data are two often observed types of non-Gaussian survey data. This class also includes the Negative-Binomial distribution, which may be used to model count data.

Specifically, Polson et al. (2013) define a random variable X to have a Pólya-Gamma distribution with parameters $b > 0$ and $c \in \mathcal{R}$, denoted $\mathrm{PG}(b, c)$, if X is equal in distribution to

$$\frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k - 1/2)^2 + c^2/(4\pi^2)},$$

where $g_k \overset{ind}{\sim} Gamma(b, 1)$. Furthermore, they show that

$$\frac{(e^{\psi})^a}{(1 + e^{\psi})^b} = 2^{-b} e^{\kappa\psi} \int_0^{\infty} e^{-\omega\psi^2/2} p(\omega) d\omega, \tag{3.3}$$

where $\kappa = a - b/2$ and $p(\omega)$ is a $\mathrm{PG}(b, 0)$ density. They also show that $(\omega|\psi) \sim \mathrm{PG}(b, \psi)$. Thus, with a Binomial likelihood, using this data augmentation scheme and

Gaussian prior distributions, one can sample from Gaussian full conditional distributions for the parameters, and Pólya-Gamma distributions for the latent augmentation variables. The `BayesLogit` package in `R` provides efficient sampling of Pólya-Gamma random variables (Windle et al., 2013)

3.2.1 Pseudo-Likelihoods

One of the main difficulties when implementing unit-level models for survey data is accounting for an informative sampling design. For example, certain demographic subgroups may be sampled with higher probability, but there may also be a relationship between these subgroups and the response variable of interest. Under this scenario, the sample is not representative of the population, and thus the sample likelihood should be adjusted to account for this. Chapter 1 gives a review of modern methods for unit-level modeling under informative sampling. One general approach is to use a pseudo-likelihood, introduced by Skinner (1989) and Binder (1983), by weighting each unit's likelihood contribution using the reported survey weight w_i,

$$\prod_{i \in \mathcal{S}} f(Z_i \mid \boldsymbol{\theta})^{w_i}, \tag{3.4}$$

where $\mathcal{S}$ indicates the sample and Z_i represents the response value for unit i.

The PL can be maximized using maximum-likelihood techniques, however Savitsky and Toth (2016) show that a PL may also be used in a Bayesian setting, thus generating a pseudo-posterior distribution

$$\hat{\pi}(\boldsymbol{\theta}|\mathbf{Z}, \tilde{\mathbf{w}}) \propto \left\{ \prod_{i \in \mathcal{S}} f(Z_i|\boldsymbol{\theta})^{\tilde{w}_i} \right\} \pi(\boldsymbol{\theta}).$$

They emphasize the importance of scaling the weights to sum to the sample size, $\tilde{w}_i = n\frac{w_i}{\sum_{j=1}^{n} w_j}$, in order to prevent contraction of the PL and achieve appropriate variance estimates.

Using a unit-level model such as this, it is simple to generate predictions for any unobserved units, thereby effectively generating the population. It is then straightforward to aggregate units in order to estimate any finite population quantities, such as for SAE purposes. Under a Bayesian framework, this can be done for each sample from the posterior distribution, thus yielding a posterior distribution over any desired estimates. In the special case where all covariates are categorical in nature, this approach can be seen as a type of poststratification (see Gelman and Little (1997) and Park et al. (2006) for examples of poststratifaction outside of a pseudo-likelihood framework). For special cases where the postratification variables include all survey design variables, poststratifaction alone may be used to account for the sample. However, this is typically not the case for complex survey designs, thus the pseudo-likelihood may be used in conjunction with poststratification. Zhang et al. (2014) provide an example of a pseudo-likelihood and poststratification combination for small area estimates in a frequentist framework, whereas Chapter 2 takes a Bayesian pseudo-likelihood and poststratification approach.

Now, an unweighted binomial likelihood has the form

$$\prod_{i \in \mathcal{S}} \frac{(e^{\psi_i})^{Z_i}}{(1+e^{\psi_i})^{n_i}}.$$

By using a pseudo-likelihood instead, the form becomes

$$\prod_{i\in\mathcal{S}}\left(\frac{(e^{\psi_i})^{Z_i}}{(1+e^{\psi_i})^{n_i}}\right)^{\tilde{w}_i}=\prod_{i\in\mathcal{S}}\frac{(e^{\psi_i})^{Z_i^*}}{(1+e^{\psi_i})^{n_i^*}}, \tag{3.5}$$

where $Z_i^* = Z_i \times \tilde{w}_i$ and $n_i^* = n_i \times \tilde{w}_i$. The PL given by (3.5) is of the same form as that given in (3.3), thus we are able to sample from conjugate full conditional distributions using a binomial type PL with Gaussian prior distributions, and PG data augmentation variables.

3.2.2 Binomial Response Model

Using the Pólya-Gamma data augmentation scheme, we develop a computationally efficient pseudo-likelihood mixed model for binomial survey data (PL-MB) under informative sampling,

$$\begin{aligned}
\boldsymbol{Z}|\boldsymbol{\beta},\boldsymbol{\eta} &\propto \prod_{i\in S}\text{Bin}\left(Z_i|n_i,p_i\right)^{\widetilde{w}_i}\\
\text{logit}(p_i) &= \boldsymbol{x_i'\beta}+\boldsymbol{\phi_i'\eta}\\
\boldsymbol{\eta}|\sigma^2_\eta &\sim \text{N}_r(\boldsymbol{0_r},\sigma^2_\eta\boldsymbol{I}_r)\\
\boldsymbol{\beta} &\sim \text{N}_q(\boldsymbol{0_q},\sigma^2_\beta\boldsymbol{I}_q)\\
\sigma^2_\eta &\sim \text{IG}(a,b)\\
\sigma_\beta, a, b &> 0,
\end{aligned} \tag{3.6}$$

where Z_i represents the response for unit $i \in \mathcal{S}$. We model the data using a Binomial pseudo-likelihood, with n_i representing the number of trials, and p_i representing the probability of a positive response (e.g. a unit having health insurance) for unit i. In many survey data scenarios, including those explored here, the data is binary, thus $n_i = 1, \forall i$. The vector $\boldsymbol{x}_i'$ represents a q-dimensional set of covariates and $\boldsymbol{\beta}$ is the q-dimensional vector of fixed effects. In this work, the vector $\boldsymbol{\phi}_i'$ represents either an r-dimensional vector of spatial basis functions, or an incidence vector, indicating which area unit i resides in. In this way, the r-dimensional vector $\boldsymbol{\eta}$ act as area-level random effects. Note that the Binomial pseudo-likelihood can be rewritten using (3.3). Although we do not present the model this way for the sake of readability, we take advantage of this fact when we construct the Gibbs sampling scheme, which introduces a step to sample the latent Pólya-Gamma random variables. The full conditional distributions for Gibbs sampling, which rely on the Pólya-Gamma data augmentation, can be found in Appendix A.2. As an alternative to Gibbs sampling, for manageable sample sizes Hamiltonian Monte Carlo could be used, for example via Stan (Stan Development Team, 2021).

3.2.3 Variational Bayes Approximation

In many high-dimensional settings, it can become a computational burden to sample from the posterior distribution via MCMC, even through the use of Gibbs sampling with fully conjugate full conditional distributions. For example, using the Pólya-Gamma data augmentation scheme, a latent random variable must be drawn for every sample observation at every iteration of the MCMC. As sample sizes become very large, this may become infeasible, even after allowing for parallel computing techniques. One popular solution to this computational problem is the variational Bayes approach (Jordan et al., 1999; Wainwright et al., 2008), for which an approximation to the posterior distribution is used rather than the true posterior distribution. A class of distributions, $\mathcal{D}$, is chosen for $q^*(\boldsymbol{\theta})$, the approximation to the true posterior, $p(\boldsymbol{\theta}|\boldsymbol{x})$. Optimization techniques may then be used to minimize the Kullback-Leibler (KL) divergence between the approximate and true posterior distributions,

$$q^*(\boldsymbol{\theta}) = \arg\min_{q(\boldsymbol{\theta})\in\mathcal{D}} \text{KL}\left(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\boldsymbol{x})\right). \tag{3.7}$$

Beal and Ghahramani (2003) focus on a specific case known as the variational Bayes EM algorithm. The approximating distribution can be factored into a product of global parameters and local latent variables, $q(\boldsymbol{\theta}) = q(\boldsymbol{\beta})\prod_{i=1}^n q(\xi_i)$. With this factorization, an iterative approach can be used to minimize the KL divergence, where

$$\begin{aligned} q(\boldsymbol{\beta})^{(t)} &\propto \exp\left\{\mathbb{E}_{q^{(t-1)}(\boldsymbol{\xi}))}\log[p(\boldsymbol{\beta}|\boldsymbol{Z},\boldsymbol{\xi})]\right\} \\ q(\xi_i)^{(t)} &\propto \exp\left\{\mathbb{E}_{q^{(t-1)}(\boldsymbol{\beta}))}\log[p(\xi_i|\boldsymbol{Z},\boldsymbol{\xi}_{-i},\boldsymbol{\beta})]\right\},\ i = 1,\ldots,n. \end{aligned} \tag{3.8}$$

In models that use fully conjugate full conditional distributions, as well as likelihoods from the exponential family, these factorized approximate distributions are of the same class as their corresponding full conditional distribution. Importantly, this includes the case of logistic regression via Pólya-Gamma data augmentation, for which Durante et al. (2019) explore a variational Bayes EM algorithm approach.

Algorithm 1 provides an extension of the one explored by Durante et al. (2019), which is intended for unweighted logistic regression, and thus not directly applicable to our pseudo-likelihood mixed model. The main extension of this algorithm is the inclusion of the pseudo-likelihood rather than the original Binomial likelihood. This algorithm may be used in place of MCMC in order to fit the PL-MB model in high dimensional settings. Independent samples from the variational approximation to the posterior of $\boldsymbol{\zeta} = (\boldsymbol{\beta'}, \boldsymbol{\eta'})$ may be drawn by sampling from a $\mathrm{N}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ distribution, which may then be used to produce any desired Monte Carlo estimates. We give more details about prediction using poststratification with the variational distribution in Appendix B.1.

Algorithm 1: VB EM algorithm for PL-MB model

```
Initialize $\tilde{\sigma}^2_\eta$ and $\tilde{\xi}_i,\ i = 1, \dots, n$ ;
Let $\boldsymbol{D} = [\boldsymbol{X}, \boldsymbol{\Phi}]$ and $\boldsymbol{\zeta} = (\boldsymbol{\beta'}, \boldsymbol{\eta'})$ ;
for $t = 1$ until convergence do
    $\tilde{\boldsymbol{\Omega}} = \mathrm{Diag}\left(\frac{w_1}{2\tilde{\xi}_1}\tanh(\tilde{\xi}_1/2), \dots, \frac{w_n}{2\tilde{\xi}_n}\tanh(\tilde{\xi}_n/2)\right)$;
    $\tilde{\boldsymbol{\Sigma}} = \left(\mathrm{blockdiag}(\frac{1}{\sigma^2_\beta}\boldsymbol{I}_p, \frac{a+r/2}{\tilde{\sigma}^2_\eta}\boldsymbol{I}_r) + \boldsymbol{D'}\tilde{\boldsymbol{\Omega}}\boldsymbol{D}\right)^{-1}$;
    $\tilde{\boldsymbol{\Sigma}}_\eta = \tilde{\boldsymbol{\Sigma}}[(p+1):(p+r), (p+1):(p+r)]$;
    $\tilde{\boldsymbol{\mu}} = (\tilde{\boldsymbol{\mu}}'_\beta, \tilde{\boldsymbol{\mu}}'_\eta)' = \tilde{\boldsymbol{\Sigma}}\boldsymbol{D'}\,(\boldsymbol{w} \odot (\boldsymbol{Z} - 1/2))$;
    $\tilde{\sigma}^2_\eta = b + \frac{1}{2}\left(\tilde{\boldsymbol{\mu}}'_\eta\tilde{\boldsymbol{\mu}}_\eta + \mathrm{tr}(\tilde{\boldsymbol{\Sigma}}_\eta)\right)$;
    for $i = 1$ to $n$ do
        $\tilde{\xi}_i = \left(\boldsymbol{D}'_i\tilde{\boldsymbol{\Sigma}}\boldsymbol{D}_i + (\boldsymbol{D}'_i\tilde{\boldsymbol{\mu}})^2\right)^{1/2}$;
    end
end
```

3.2.4 Multinomial Response Model

In addition to Binomial data, Multinomial or categorical data is often observed in survey data. In a similar fashion as the PL-MB model, we can write the Pseudo-likelihood mixed effect Multinomial model (PL-MM) with K categories as

$$\begin{aligned}
\boldsymbol{Z}|\boldsymbol{\beta},\boldsymbol{\eta} &\propto \prod_{i\in S} \text{Multinomial}\left(\boldsymbol{Z_i}|n_i,\boldsymbol{p_i}\right)^{\widetilde{w}_i} \\
p_{ik} &= \frac{\exp(\psi_{ik})}{\sum_{k=1}^{K}\exp(\psi_{ik})} \\
\psi_{ik} &= \boldsymbol{x_i'\beta_k} + \boldsymbol{\phi_i'\eta_k} \\
\boldsymbol{\eta_k}|\sigma^2_{\eta k} &\sim \text{N}_r(\boldsymbol{0_r},\sigma^2_{\eta k}\boldsymbol{I}_r),\ k=1,\dots,K-1 \\
\boldsymbol{\beta_k} &\sim \text{N}_p(\boldsymbol{0_p},\sigma^2_{\beta}\boldsymbol{I}_p),\ k=1,\dots,K-1 \\
\sigma^2_{\eta k} &\sim \text{IG}(a,b),\ k=1,\dots,K-1 \\
&\sigma_\beta, a, b > 0,
\end{aligned} \tag{3.9}$$

where $\boldsymbol{\beta_K}$ and $\boldsymbol{\eta_K}$ are constrained to be equal to zero for identifiability. The K-dimensional vector $\boldsymbol{Z_i}$ represents the number of successful outcomes in each of the K categories for survey unit i, and the K-dimensional vector $\boldsymbol{p}_i$ represents the probability of each category for unit i.

Although Algorithm 1 is intended for Binomial data, a stick-breaking representation of the Multinomial distribution can be used to expand the applicability of this VB approach. Specifically, Linderman et al. (2015) show that the Multinomial distribution may be written as a product of independent Binomial distributions,

$$\text{Multinomial}(\boldsymbol{Z}|n,\boldsymbol{p}) = \prod_{k=1}^{K-1}\text{Bin}(Z_k|n_k,\tilde{p}_k), \tag{3.10}$$

where

$$n_k = n - \sum_{j<k} Z_j,\quad \tilde{p}_k = \frac{p_k}{1-\sum_{j<k}p_j},\quad k=2,\dots,K. \tag{3.11}$$

Under this view of Multinomial data, we can rewrite the PL-MM model as

$$\begin{aligned}
\boldsymbol{Z}|\boldsymbol{\beta},\boldsymbol{\eta} &\propto \prod_{i\in S}\prod_{k=1}^{K-1} \text{Bin}\left(Z_{ik}|n_{ik},\tilde{p}_{ik}\right)^{\widetilde{w}_i} \\
\text{logit}(\tilde{p}_{ik}) &= \boldsymbol{x_i'\beta_k} + \boldsymbol{\phi_i'\eta_k} \\
\boldsymbol{\eta_k}|\sigma^2_{\eta k} &\sim \text{N}_r(\boldsymbol{0_r},\sigma^2_{\eta k}\boldsymbol{I}_r),\ k=1,\ldots,K-1 \\
\boldsymbol{\beta_k} &\sim \text{N}_p(\boldsymbol{0_p},\sigma^2_{\beta}\boldsymbol{I}_p),\ k=1,\ldots,K-1 \\
\sigma^2_{\eta k} &\sim \text{IG}(a,b),\ k=1,\ldots,K-1 \\
\sigma_\beta, a, b &> 0,
\end{aligned} \tag{3.12}$$

where $n_{ik} = n_i - \sum_{j<k} Z_{ij}$ and $\tilde{p}_{ik} = \frac{p_{ik}}{1-\sum_{j<k} p_{ij}}, \quad k = 2,\ldots,K$. Thus, the PL-MM model may be fit as a series of $K-1$ independent Binomial models using either MCMC or the VB approach outlined in Algorithm 1. Note that after fitting the model, the stick breaking probabilities $\tilde{\boldsymbol{p}}_i$ can be transformed back to the original probabilities $\boldsymbol{p}_i$ for inference.

3.3 Empirical Simulation Study

In order to mimic a real survey data setting, our simulations revolve around resampling of an existing survey dataset rather than generating a synthetic population from a parametric distribution. Specifically, we treat the existing survey sample as our population and then take a further sample with probability proportional to s_i, a size variable that is constructed in an informative manner. This informative sampling scheme can be validated by comparing the weighted design-based estimator to an unweighted design-based estimator. Under an informative design, the unweighted estimator will result in greater bias.

3.3.1 Multinomial Response Simulation

An important SAE application is the Small Area Health Insurance Estimation (SAHIE) program (Bauder et al., 2018). The goal of SAHIE is to estimate the proportion of individuals with health insurance by county for a number of income to poverty ratio (IPR) categories. IPR is defined as family income, divided by the appropriate federal poverty level. The IPR categories under consideration are 0-138%, 138-200%, 200-250%, 250-400%, and 400+%. The thresholds for the first three IPR categories are motivated, in part, by needs of the Centers for Disease Control and Prevention, which provides breast and cervical cancer screenings for low income and uninsured women. The IPR categories are also relevant to the Affordable Care Act, which increased access to health insurance. In participating states, Medicaid programs expanded health insurance access to individuals and families with IPR less than 138%, and provided tax credits for those with IPR between 138% and 400%.

The true number of people within each IPR category is unknown and must be estimated. Thus, to create estimates of the proportion with health insurance by IPR category, health insurance and IPR category must be modeled simultaneously. Within each IPR category, an individual may be categorized as either having or not having health insurance. In this manner, we view individuals as falling into one of 10 distinct categories, $(C_{1,0}, \ldots, C_{5,0}, C_{1,1}, \ldots, C_{5,1})$, where $C_{j,k}$ indicates an individual in IPR category $j = 1, \ldots, 5$ and health insurance indicator $k = 0, 1$. The multivariate structure of the data, with many IPR categories, and the need to estimate both the number in each IPR category along with the proportion with health insurance, makes unit-level modeling appealing for this dataset. In addition, there are areas for which there is no sample, areas for which there are direct estimates on the boundary of the parameter space that are exactly equal to zero or one, and direct estimates for which the sampling variance estimates are not well defined, which makes area-level modeling challenging. Since there is no established theory for applying area-level methodology to this type of survey data, we restrict our simulation study to unit-level methods.

To construct health insurance estimates by county and IPR category, we fit the PL-MM with ten categories using $n_i = 1$ for all i. We let $\boldsymbol{x}_i$ consist of poststratification variables including race category, sex, and age category. We also let $\boldsymbol{\phi}_i$ be a vector indicating which county unit i resides in. Thus, the model uses a county level random effect. We use a vague prior distribution over $\boldsymbol{\beta}$ and σ_η^2 by setting $\sigma_\beta^2 = 1000$ and $a = b = 0.5$. A sensitivity analysis confirmed that these prior choices had very little effect on the model outcome, but for other data scenarios, this choice should be considered carefully. The model is fit using both the MCMC and VB fitting strategies, with both drawing a posterior sample size of 1000, after discarding 1000 draws as burnin for MCMC. For MCMC, convergence was assessed visually through the use of traceplots of the sample chains along with the Geweke convergence diagnostic (Geweke, 1991), for which no lack of convergence was detected. After fitting the model on the sample data, predictions are made for all units in the population. The synthesized population is then aggregated to the desired level of the estimates (i.e., county by IPR category). This is done for each posterior draw, giving a posterior predictive distribution for the desired estimates.

To assess the SAE capability of our PL-MM model through simulation, we treat the 2014 1-year American Community Survey (ACS) sample in Minnesota as our population. This data contains roughly 120,000 respondents across Minnesota's 87 counties. We then take a further probability proportional to size sample without replacement, using the Poisson method (Brewer et al., 1984) with an expected sample size of 10,000. We use the size variable $s_i = \exp\{w_i^* + 2\text{I}(H_i = 0)\}$, where w_i^* is the original survey weight for unit i after scaling to have mean zero and standard deviation of one, and H_i indicates whether or not unit i had health insurance. Estimates are constructed using the PL-MM with both MCMC and VB fits. We also construct a Horvitz-Thompson direct estimate as well as an unweighted direct estimate. We repeat the sampling and estimation process 50 times in order to compare MSE and bias across estimators.

A summary of the simulation results in given in Table 3.1, including average mean squared error (MSE) and squared bias for the competing estimators, as well as computation time and 95% credible interval (CI) coverage rates for the two model based estimators. The higher bias of the UW estimator relative to the direct estimator indicates that the sampling scheme was indeed informative. The two model based approaches yield significant reductions to MSE when compared to the direct estimator. Surprisingly, the predictions using the VB approach had even lower MSE than the predictions using MCMC. The reason for the reduced MSE is not entirely clear and is a subject for future research. The downside to the VB approach is that the approximate posterior results in uncertainty estimates that are not optimal. This is reflected in the lower 95% CI coverage rate for the VB approach compared to the MCMC approach. This is to be expected, as the VB approach only approximates the true posterior distribution. However, the differences are relatively minor, and can be justified through the massive decrease in computation time.

Estimator	MSE	Bias^2	Time (s)	Coverage Rate
Model MCMC	7.1×10^{-3}	3.7×10^{-3}	7314	**94%**
Model VB	$\mathbf{2.3 \times 10^{-3}}$	$\mathbf{1.7 \times 10^{-3}}$	**140**	87%
Direct	9.9×10^{-2}	3.8×10^{-2}	-	-
UW Direct	1.6×10^{-1}	1.1×10^{-1}	-	-

Table 3.1: Simulation results: MSE and squared bias of the four estimators averaged across counties. Average computation time in seconds and 95% credible interval coverage rate are also given for the model based estimates.

We also show the MSE by county and IPR category for each estimator in Figure 3.1. The largest reductions in MSE through model-based estimation tend to occur for the more rural and sparsely populated regions of the state. These counties tend to have smaller sample sizes resulting in more erratic direct estimates. The model-based estimates borrow strength from sampled units in all counties, resulting in more stable (i.e. lower MSE) estimates.

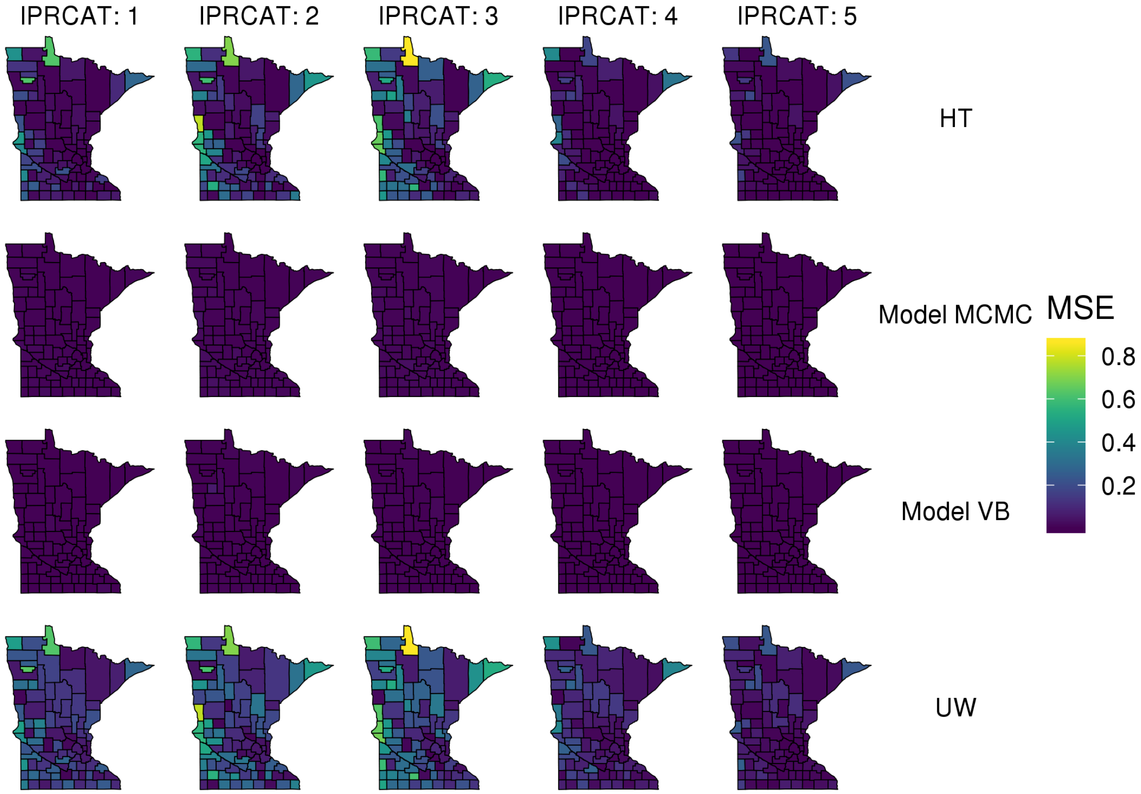

Figure 3.1: Emprical mean squared error by county across the simulation based estimates for the state of Minnesota. Columns represent the different IPR categories and rows represent the different estimators.

3.4 Data Analysis

The simulation in Section 3.3 illustrates how the PL-MM model may be used to generate SAHIE type estimates for a single state. However, the SAHIE program is tasked with creating estimates for the entirety of the U.S. rather than a single state. The bottleneck in the MCMC approach to the PL-MM model is the generation of Pólya-Gamma random variables for every sample observation at every MCMC iteration. Although this approach is feasible at a state level, it becomes unwieldy at the national level, where the ACS samples 3.5 million households annually. For this reason, we rely on the VB approach to the PL-MM model in order to create estimates of health insurance by county and IPR category for the entire continental U.S.

Again, we use the PL-MM model with 10 categories and $n_i = 1$ for all i. We also use the same prior distribution and poststratification variables that were considered in Section 3.3. There are over 3,000 counties in the US, compared to only 87 in Minnesota, thus, we require a form of dimension reduction for ϕ_i rather than using county indicators. To do this, we let ϕ_i be equal to a set of spatial basis functions evaluated for unit i. Specifically, for illustration, we use the first 307 (10%) eigenvectors of the county adjacency matrix as our spatial basis functions. This choice was motivated in part by the suggestion of Hughes and Haran (2013) to use 10% of the available eigenvectors, as well as by the need for substantial dimension reduction with respect to the random effects. In this problem, due to modeling at the unit-level, some form of dimension reduction is needed to avoid having memory issues that would result from an approximately 4.5 million × 3000 dimensional matrix. A limited visual sensitivity analysis was used in the selection of basis functions here, however, before a model such as this is implemented in a production setting, it may be worthwhile to conduct a more comprehensive sensitivity analysis. Choosing the number of basis functions is problem specific and constitutes an ongoing area of research; e.g., see Bradley et al. (2016) and the references therein.

We fit the PL-MM model using the VB approach, with a sample size of roughly 4.5 million. We then take 1000 independent draws from the variational posterior distribution in order to construct the posterior predictive distribution of our estimates.

Treating the posterior predictive mean as our point estimates, we plot the model-based estimates alongside the direct estimates in Figure 3.2. In order to satisfy the disclosure avoidance requirements of the U. S. Census Bureau, a small amount of noise was added to the direct estimates shown in the maps in Figure 3.2. However, this is the only instance of any additional noise being added to data or estimates. All models were fit to the raw ACS data, and all other results are presented without any additional noise. Note that there is an unpopulated county in Wyoming. This county is shown as white due to the fact that there are no units to aggregate during the poststratification procedure (see Appendix B.1 for details on poststratification). Visually, the direct estimates are quite noisy, due to the very small sample sizes in many counties. The model based estimates are able to provide a degree of smoothing through the use of borrowed information in the hierarchical model structure. This results in model based estimates that have the same general spatial pattern as the direct estimates without as much noise. We also plot the health insurance estimates by county without regard to IPR category in Figure 3.3. Similar patterns can be noticed here.

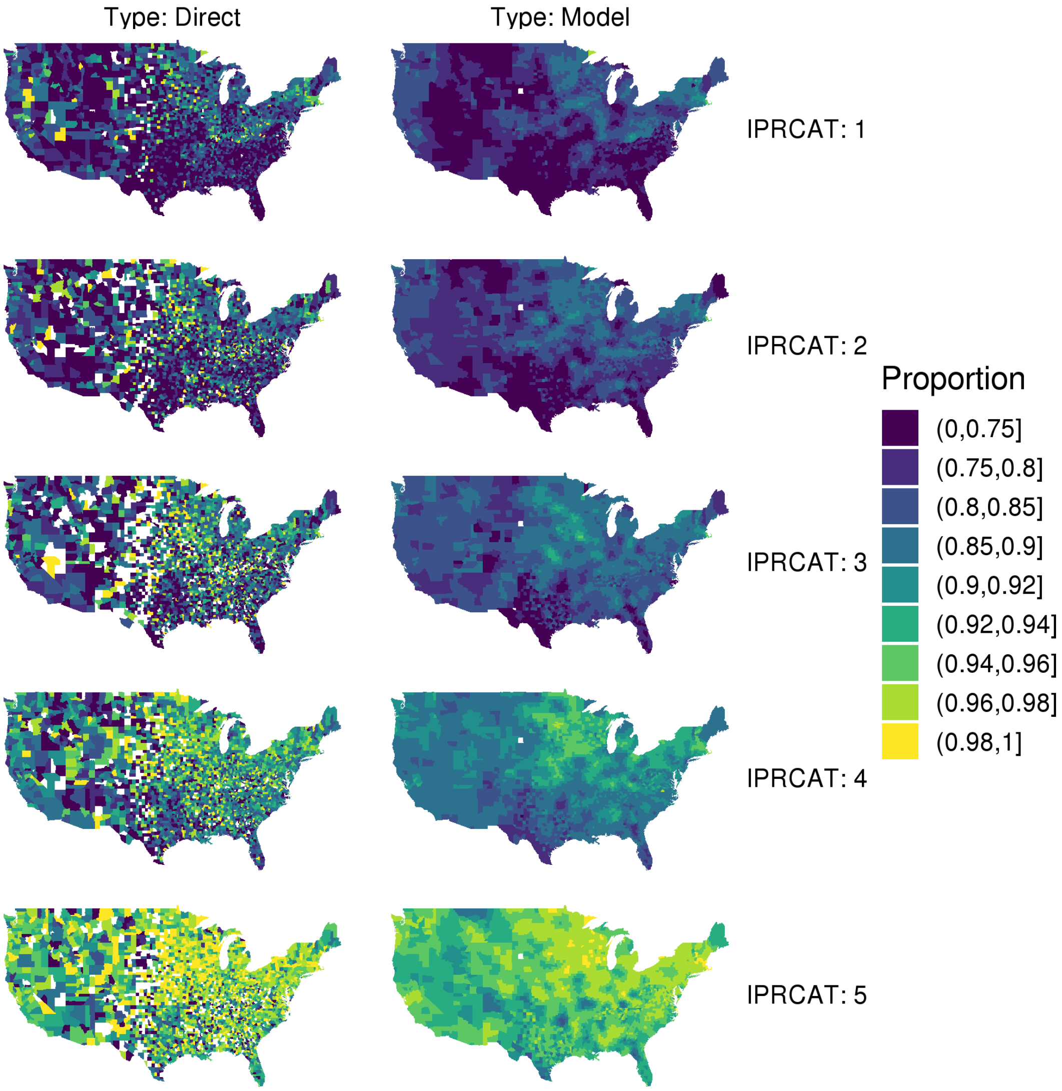

Figure 3.2: Direct and model based estimates of the proportion of the population with health insurance by county and IPR category for the continental United States.

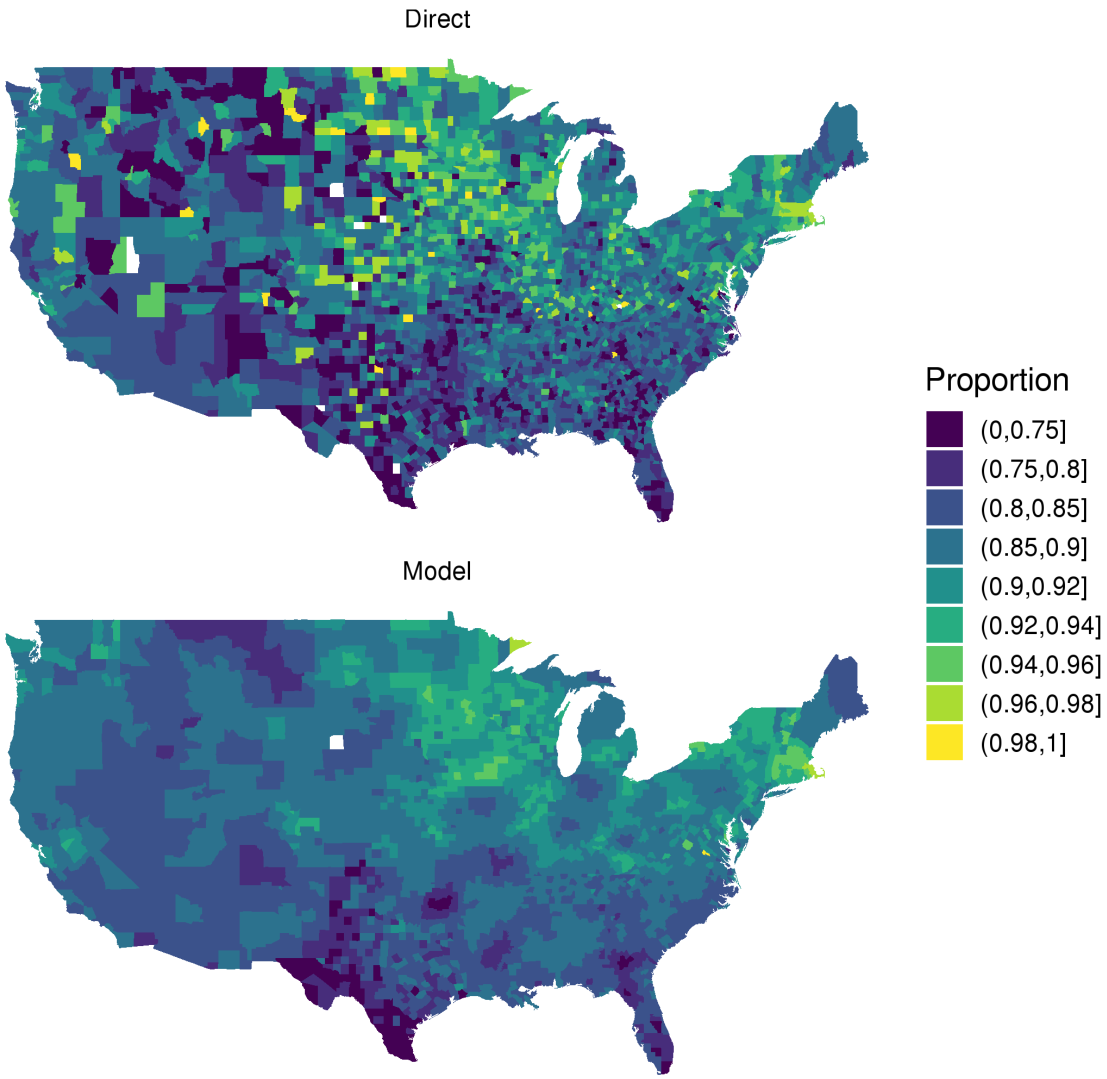

Figure 3.3: Direct and model based estimates of the proportion of the population with health insurance by county for the continental United States.

We plot the ratio of the model based standard errors to the direct estimate standard errors by county and IPR category in Figure 3.4. For the vast majority of estimates, the model based approach provides quite substantial reductions in standard error, with the largest advantage occurring in the more sparsely populated Southern and Western regions of the country.

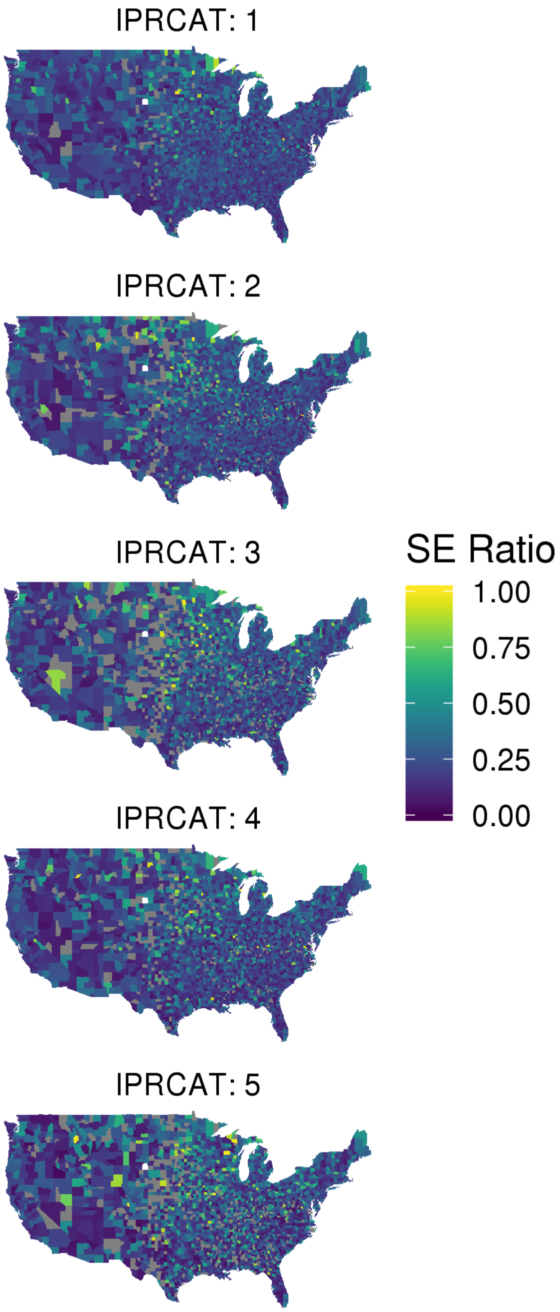

Figure 3.4: Ratio of model based standard errors to direct estimate standard errors by county and IPR category for the continental United States. Counties with no available direct estimate are shown in gray.

This example demonstrates how the PL-MB and PL-MM models may be used to model complex dependence structures with non-Gaussian data in a computationally efficient manner. The VB approach specifically was able to generate estimates for over 15,000 county and IPR category combinations, utilizing a sample size of over 4 million, in roughly 17 hours. These estimates are much less noisy than direct estimates, with substantially lower standard errors. Furthermore, the simulation results of Section 3.3 indicate that these model based estimates should have much lower MSE. In addition to advantages over the direct estimate, this approach has many advantages over area-level modeling approaches, such as the one currently in use for SAHIE. For example, unit-level models allow for easy aggregation to multiple domains. A single PL-MM model may be used to give county and state level estimates, whereas area-level modeling strategies require two separate models and often rely on ad-hoc benchmarking techniques. Another advantage is that unit-level models do not require a direct estimate for a given area in order to construct an estimate, in contrast to area-level models.

3.5 Discussion

This chapter establishes a framework for modeling Binomial and Multinomial unit-level survey data, specifically under an informative sample. We envision this methodology being used to create area-level estimates of population proportions, with health insurance (SAHIE) as our motivating example. The current methodology used to generate SAHIE estimates is conducted at the area level which can cause a number of problems that are alleviated through the use of unit-level modeling. Our unit-level approach is able to generate multiple levels of estimates through a single model without the need for benchmarking techniques. We demonstrate this by producing health insurance estimates by county as well as by IPR category within each county for the entire continental U.S. Our approach is also able to produce very precise estimates compared to traditional direct estimators, as demonstrated by our empirical simulation study. Finally, these estimates can be produced in a very computationally efficient manner either through the use of either Gibbs sampling with fully conjugate full-conditional distributions or through a VB approximation to the posterior distribution.

Although this chapter provides a methodological step forward for small area estimates of health insurance, further work would be necessary to create estimates that might replace the current SAHIE program. For example, the current SAHIE methodology considers a number of important covariates that were not considered here, due to disclosure limitations, including data from the Supplemental Nutrition Assistance Program as well as Medicaid. Furthermore, the method considered here is a type of generalized linear model, but there is potential for improvement through the use nonlinear modeling techniques (e.g. see Chapter 4).

Chapter 4

Computationally Efficient Deep Bayesian Unit-Level Modeling of Survey Data under Informative Sampling for Small Area Estimation

4.1 Introduction

There has recently been a strong interest in collecting novel types of data along with sample surveys. These data types can range from functional data such as the physical activity monitor data contained within the National Health and Nutrition Examination Survey (NHANES; Schuna et al., 2013), free response text data such as those within the American National Election Studies (ANES) surveys (DeBell, 2013), and even complex data from web-based surveys such as mouse movements

(Horwitz et al., 2020). These information rich data sources could prove to be useful as covariates, however the rapid and increased interest in these data types within a survey context has resulted in lagging development of corresponding methodology.

These complex covariates are generally collected at the unit level, and thus, necessitate the need for unit-level modeling strategies. One of the challenges associated with unit-level modeling is accounting for the sampling design under informative sampling mechanisms. A variety of strategies exist for this problem, including the use of pseudo-likelihood modeling (Skinner, 1989; Binder, 1983). In many applications, such as small area estimation, predictions can be made for every unit in the population and then aggregated as necessary to construct any desired estimates. An exceedingly common solution is that of regression and poststratification, whereby the population is segmented via a set of categorical covariates, and units that are associated with identical covariates are assumed to be independent and identically distributed (Park et al., 2004). In particular, Park et al. (2004) state that they envision this methodology being used for public opinion estimates at the state level. The categorical covariates required for poststratification are generally known for the entire population. This is in contrast to the complex covariates that we consider, which are generally only known for the sampled units.

Another major concern for unit-level models, particularly in a Bayesian setting, is that of computational efficiency. Dependent data models typically rely on Gaussian prior distributions for model parameters (Bradley et al., 2015); however, most survey variables tend to be non-Gaussian, leading to non-conjugate conditional distributions that can be difficult to sample from. This problem is addressed by Chapter 2 and Chapter 3 for count data and Binomial data, although they do not consider the further problem of modeling complex data types in a computationally efficient manner.

Herein, we develop a computationally efficient method to model these complex covariates while accounting for informative sampling. Although other applications of this model are possible, we illustrate this method through the problem of small area estimation. We utilize a neural network structure to handle nonlinear modeling of the complex covariate data, while employing a Bayesian pseudo-likelihood model structure in order to measure uncertainty around our estimates while accounting for informative sampling. The remainder of this chapter is outlined as follows. In Section 4.2 we introduce the necessary methodological background as well as our model. Section 4.3 considers an empirical simulation study relying on the use of ANES data. We follow with a full data analysis using the same ANES data in Section 4.4. Finally, we provide discussion and concluding remarks in Section 4.5.

4.2 Methodology

The method that we develop tackles three problems simultaneously. The first issue is that nonlinear modeling is required for the use of complex covariates, without becoming computationally prohibitive. For example, neural network structures typically involve an extremely high-dimensional parameter space. A full Bayesian treatment of these types of models can often require too many computational resources to be fit in a reasonable amount of time. The second problem is that we must account for informative sampling in our model to avoid producing any unnecessary statistical bias. Finally, we require a model that can handle Binomial data types while still accounting for all of the underlying dependencies associated with the data. We explore each of these three problems, and then present our methodology.

4.2.1 Extreme Learning Machines

The extreme learning machine (ELM) is a type of single layer feed-forward neural network (FNN), introduced by Huang et al. (2006). The key difference between the ELM and traditional FNNs is that the ELM uses random weights (i.e., parameters) drawn from some distribution for the hidden layer nodes. As with other FNNs, the ELM can be used for both regression and classification problems, as well as other types of problems (e.g., unsupervised learning), while allowing for much more flexibility in the mean function than linear or generalized linear models.

The basic ELM considers a nonlinear transformation of the covariate data (features),

$$f(\mathbf{x}_i) = \sum_{j=1}^{N} g_j(\mathbf{a}_j'\mathbf{x}_i + b_j)\beta_j, \ i = 1, \ldots, n$$

where $\mathbf{x}_i$ represents the p-dimensional covariate information for unit i in the sample with size n. The value N represents the number of nodes, where each node considers a unique nonlinear transformation of the data. Each node first applies a linear transformation, with parameters $\mathbf{a}_j = [a_{j1}, \ldots, a_{jp}]'$ and b_j. This linear transformation is followed by a nonlinear transformation, denoted by the function $g_j(\cdot)$, often called an activation function. This is specified a priori, and may be any piecewise continuous function (Huang et al., 2015), but in practice usually consists of a sigmoid function. The output, $f(\mathbf{x}_i)$ is then calculated as a weighted sum of each individual node output, where the weights are denoted by the N dimensional vector $\boldsymbol{\beta}$. Intuitively, this is similar to basis function approaches, in the sense that both techniques may use nonlinear transformations of the input data to construct a flexible nonlinear function for the mean. However, one key difference is that in the case of the ELM, these nonlinear transformations do not need to be selected as they are randomly generated.

The key to ELMs is that for each node, $j = 1, \ldots, N$, the values of $\mathbf{a}_j$ and b_j are randomly drawn. Thus, the only set of parameters that need to be learned or estimated is $\boldsymbol{\beta}$. Common distributional choices for these randomly selected parameters are Normal(0,1) and Uniform(-1,1). Although these hidden layer parameters are only randomly drawn a single time, typically many nodes are used to allow for flexible representation of the function $f(\mathbf{x}_i)$.

More generally, the ELM can be written

$$\begin{aligned}\boldsymbol{\mu}_i &= g_o(\mathbf{B}\mathbf{g}_i)\\ \mathbf{g}_i &= g(\mathbf{A}\mathbf{x}_i)\end{aligned}$$

where the $p \times 1$ dimensional vector $\mathbf{x}_i$ now contains an intercept and the l-dimensional vector of means, $\boldsymbol{\mu}_i$, can now incorporate multivariate responses. Also, $\mathbf{A}$ is an $N \times (p+1)$ dimensional matrix of hidden layer weights, and $\mathbf{B}$ is an $l \times N$ dimensional vector of output weights. The hidden layer activation function is denoted $g(\cdot)$ and the output layer activation function is denoted $g_o(\cdot)$, which will be the inverse of the canonical link function in the case of a GLM.

The above view of the ELM is similar to the generalized linear model and highlights an important strength of ELM. Because the hidden layer parameters are randomly chosen and not estimated, we may view the hidden layer transformations as fixed once these parameters have been generated. This is similar to regression with basis expansions as is often seen when using generalized additive models (GAMs). A key difference with ELM compared to GAMs however, is that the entire vector $\mathbf{x}_i$ is used within each hidden node, which allows for interaction effects. Viewing the random transformations as fixed allows us to extend the entire class of generalized linear models to incorporate nonlinear behavior. Furthermore, pseudo-likelihood approaches may be used in conjunction with the ELM in order to account for informative sampling.

The ELM can be considered a type of reservoir computing (an approach where weights are randomly generated). Random projection is another type of reservoir computing, often used for dimension reduction (Bingham and Mannila, 2001). Under random projection, the original $n \times p$ data matrix $\mathbf{X}$ is "projected" onto a L-dimensional subspace, $\mathbf{X}^* = \mathbf{X}\mathbf{R}$, where $\mathbf{R}$ is a randomly generated $p \times L$ matrix. This is not technically a projection, as the matrix $\mathbf{R}$ is not orthogonal, but due to the random nature of the matrix, it tends to be "approximately" orthogonal. Note that the randomly generated projection matrix could be orthogonalized, but this is not always done in practice due to the computational cost.

Random projection could be used in the context of regression, similar to Principal Components Regression. In this light, it may be seen as a special case of the ELM, where $g_j(\cdot)$ is equal to the identity function for all j. In other words, random projection uses randomly generated parameters for the hidden node linear transformation component, but does not introduce a nonlinear component.

Another common type of reservoir computing is known as the Echo State Network or ESN (Prokhorov, 2005). This is a type of recurrent neural network, where the hidden weights are randomly generated. Recently, the ESN has been used in likelihood-based frameworks for spatio-temporal forecasting (McDermott and Wikle, 2017). The ESN may also be used within a Bayesian model structure in order to give uncertainty quantification (McDermott and Wikle, 2019).

Bayesian methods have been considered in the ELM community as well, beginning with Soria-Olivas et al. (2011). They consider ridge regression fit with an Empirical Bayes procedure. This achieves both regularization as well as uncertainty quantification for the output layer weights and data model variance. They also show that this method tends to give better out of sample predictions compared to the traditional ELM. Chen et al. (2016) use a variational Bayes approach to fit a Bayesian ELM. By doing so, it is possible to reduce the computational burden of the Bayesian ELM substantially.

4.2.2 Pseudo-likelihood based SAE

When fitting models with unit-level survey data, it may be the case that there exists a dependence relationship between the unit probabilities of selection, and the response values. This is termed *informative sampling*, and if this relationship is not accounted for, any estimates may be biased (Pfeffermann and Sverchkov, 2007). A thorough review of the modern approaches to handling informative sampling is given in Chapter 1. One popular approach to this problem is the use of a pseudo-likelihood (PL), introduced by Skinner (1989) and Binder (1983). The general idea is to use the reported survey weights to exponentially re-weight the likelihood contribution of each survey unit. Thus, the PL is written as

$$\prod_{i \in \mathcal{S}} f(y_i \mid \boldsymbol{\theta})^{w_i}, \tag{4.1}$$

where y_i is the response value and w_i is the survey weight for unit i in the sample $\mathcal{S}$. The PL can be maximized in order to make frequentist inference, however Savitsky and Toth (2016) show that in a Bayesian setting, the pseudo-posterior distribution,

$$\hat{\pi}(\boldsymbol{\theta}|\mathbf{y}, \tilde{\mathbf{w}}) \propto \left\{ \prod_{i \in \mathcal{S}} f(y_i|\boldsymbol{\theta})^{\tilde{w}_i} \right\} \pi(\boldsymbol{\theta}),$$

converges to the population posterior distribution, justifying the use of a Bayesian PL for inference on nonsampled units. In this case, $\tilde{w}_i$ represents the survey weights after scaling to sum to the sample size in order to give proper uncertainty quantification.

4.2.3 Logistic Models

Many survey data variables tend to be non-Gaussian at the unit level. For example, the American Community Survey contains a binary indicator of health insurance status as well as many categorical variables such as primary language spoken. In regression frameworks with non-Gaussian responses and Normal prior distributions on any regression parameters, non-conjugate full conditional distributions arise. This may lead to the need for Metropolis steps within the MCMC routine that can be prohibitively difficult to tune.

For the case of logistic models (Binomial, Negative Binomial and Multinomial responses), Polson et al. (2013) introduce a data augmentation scheme that gives rise to conjugate full-conditional distributions. This strategy relies on the use of Pólya-Gamma (PG) random variables. Specifically, they rewrite the Binomial likelihood as,

$$\frac{(e^{\psi})^a}{(1+e^{\psi})^b} = 2^{-b} e^{\kappa\psi} \int_0^{\infty} e^{-\omega\psi^2/2} p(\omega) d\omega, \tag{4.2}$$

where $\kappa = a - b/2$ and $p(\omega)$ is a PG$(b, 0)$ density. They also show that $p(\omega|\psi) \sim$ PG(b, ψ). For the linear predictor $\psi = \boldsymbol{x}'\boldsymbol{\beta}$, if we use a Gaussian prior on $\boldsymbol{\beta}$, the full conditional distribution for $\boldsymbol{\beta}$ will also be Gaussian. Furthermore, Chapter 3 shows that under a PL setup, conjugacy is still retained. It develops both a Gibbs sampling algorithm as well as a variational Bayes algorithm for PL-based mixed effects models with Binomial data. In addition to this, it uses the stick-breaking representation of the Multinomial distribution in order to extend the algorithms to categorical responses. More specifically, Linderman et al. (2015) show that the Multinomial distribution may be written as a product of independent Binomial distributions,

$$\text{Multinomial}(\boldsymbol{Z}|n, \boldsymbol{p}) = \prod_{k=1}^{K-1} \text{Bin}(Z_k|n_k, \tilde{p}_k), \tag{4.3}$$

where

$$n_k = n - \sum_{j<k} x_j, \quad \tilde{p}_k = \frac{p_k}{1 - \sum_{j<k} p_j}, \quad k = 2, \ldots, K. \tag{4.4}$$

Under this view of Multinomial data, $K - 1$ Binomial data models may be fit independently while still accounting for the dependence between categories through the stick-breaking counts and probabilities.

4.2.4 Proposed Model

We now introduce a Bayesian unit-level deep model for informative sampling (BUDIS). Here, we focus on the case of Binomial and Multinomial data, but note that this approach would be applicable to Gaussian data as well. The Binomial model is written,

$$
\begin{aligned}
\boldsymbol{Z}|\boldsymbol{\beta},\boldsymbol{\eta} &\propto \prod_{i\in S}\left\{\text{Bin}\left(Z_i|n_i,p_i\right)^{\tilde{w}_i}\right\} \\
\text{logit}(p_i) &= \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{g}_i'\boldsymbol{\eta} \\
\boldsymbol{g_i'} &= \frac{1}{1+e^{-\boldsymbol{A}\boldsymbol{\psi_i}}} \\
\boldsymbol{\eta}|\sigma_\eta^2 &\sim \text{N}_h(\boldsymbol{0}_h, \sigma_\eta^2\boldsymbol{I}_h) \\
\boldsymbol{\beta} &\sim \text{N}_p(\boldsymbol{0}_p, \sigma_\beta^2\boldsymbol{I}_p) \\
\sigma_\eta^2 &\sim \text{IG}(a,b) \\
a, b, \sigma_\beta^2 &> 0,
\end{aligned} \tag{4.5}
$$

where Z_i is the Binomial response for unit i in the sample with size n_i and probability p_i. Typically surveys contain Bernoulli data, so for our purposes, $n_i = 1$ for all i. We are using a pseudo-likelihood approach at this data stage of the model in order to account for informative sampling. Within the pseudo-likelihood, we use the scaled survey weights, $\tilde{w}_i$, such that the weight sum to the sample size. The length p vector $\boldsymbol{x}_i$ contains any covariates that do not require nonlinear modeling. The length h vector $\boldsymbol{g}_i$ contains the ELM hidden layer values for unit i. Finally, the length r vector $\boldsymbol{\psi}_i$ contains the complex covariates that are used within the ELM framework. Note that the $h \times r$ matrix $\boldsymbol{A}$ is sampled and considered fixed before model fitting, so that $\boldsymbol{g}_i$ is determined *a priori*. This allows for the use of generlized linear model fitting procedures rather than custom techniques. Specifically, we use the variational Bayes procedure from Chapter 3 for all model fitting. We note that for this approach to truly be considered deep learning, the ELM component of the model would require multiple hidden layers (see Chamara et al. (2013)). In our case, we did not find any benefit from the inclusion of multiple layers, however, for other applications, it is straighforward to extend the ELM to multiple hidden layers.

For our purposes, we let $\boldsymbol{x}_i$ consist of any poststratification variables as well as spatial basis functions. These values will typically be known for the full population. The complex covariates contained within $\boldsymbol{\psi}_i$ are not usually known for the full population, and thus must be imputed in order to generate the population posterior predictive distribution necessary for small area estimation. Our approach revolves around the idea of assigning the observed covariate vectors to all the unobserved population units. Under a simple random sample, a reasonable assumption may be that the observed complex covariates are uniformly distributed throughout the population. However, under an informative sample, the observed covariate vectors are sampled with unequal probability, which should be accounted for when distributing the observed vectors to the population.

For our imputation model, we create imputation cells, similar to poststratification cells, with J total cells. A population unit within imputation cell $j,\ j = 1, \ldots, J$ may only be assigned a vector from the set of observed vectors $(\boldsymbol{\psi}_{j1}, \ldots, \boldsymbol{\psi}_{jn_j})$, where n_j is the sample size within cell j. Rather than sampling from this set with equal probability, we sample with probability proportional to the reported sampling probability, or inversely proportional to the reported sample weight, to account for the original survey sampling scheme. Thus, for population unit i in cell j, the vector of complex covariates is sampled from $(\boldsymbol{\psi}_{j1}, \ldots, \boldsymbol{\psi}_{jn_j})$ with probability proportional to $(1/w_{j1}, \ldots, 1/w_{jn_j})$. This imputation can be done a single time, however we opt to create a separate imputed dataset for each sample from our model based posterior distribution in order to account for the imputation uncertainty within our posterior predictive distribution. For this work, we let the $J = 48$ corresponding to the states where area level estimates are made, however other choices of imputation cells could be explored. One limitation to this approach is that all imputation cells must have at least one sample unit in order to distribute the sample values within the cell to the population.

4.3 Empirical Simulation Study

To test our methodology, we consider data from the 2012 American National Election Studies (ANES) survey. Specifically, we use the Time Series Study data which measures various responses both pre and post election. We only consider the post election data, which contains a number of free response questions. Our goal is to use the free responses from the question "What are the most important problems facing this country?" in order to improve small area estimates of public opinion.

Figure 4.1 shows a word cloud of the most frequently occurring words within the ANES data. In many cases, the words will have little meaning on their own, but instead have meaning when paired with other words. For example, the word *security* on its own does not provide much insight, but when paired with either *economic* or *national*, it may indicate the primary concern of the respondent. This suggests the need for a model that can take into account many possible interactions between words rather than considering words individually.

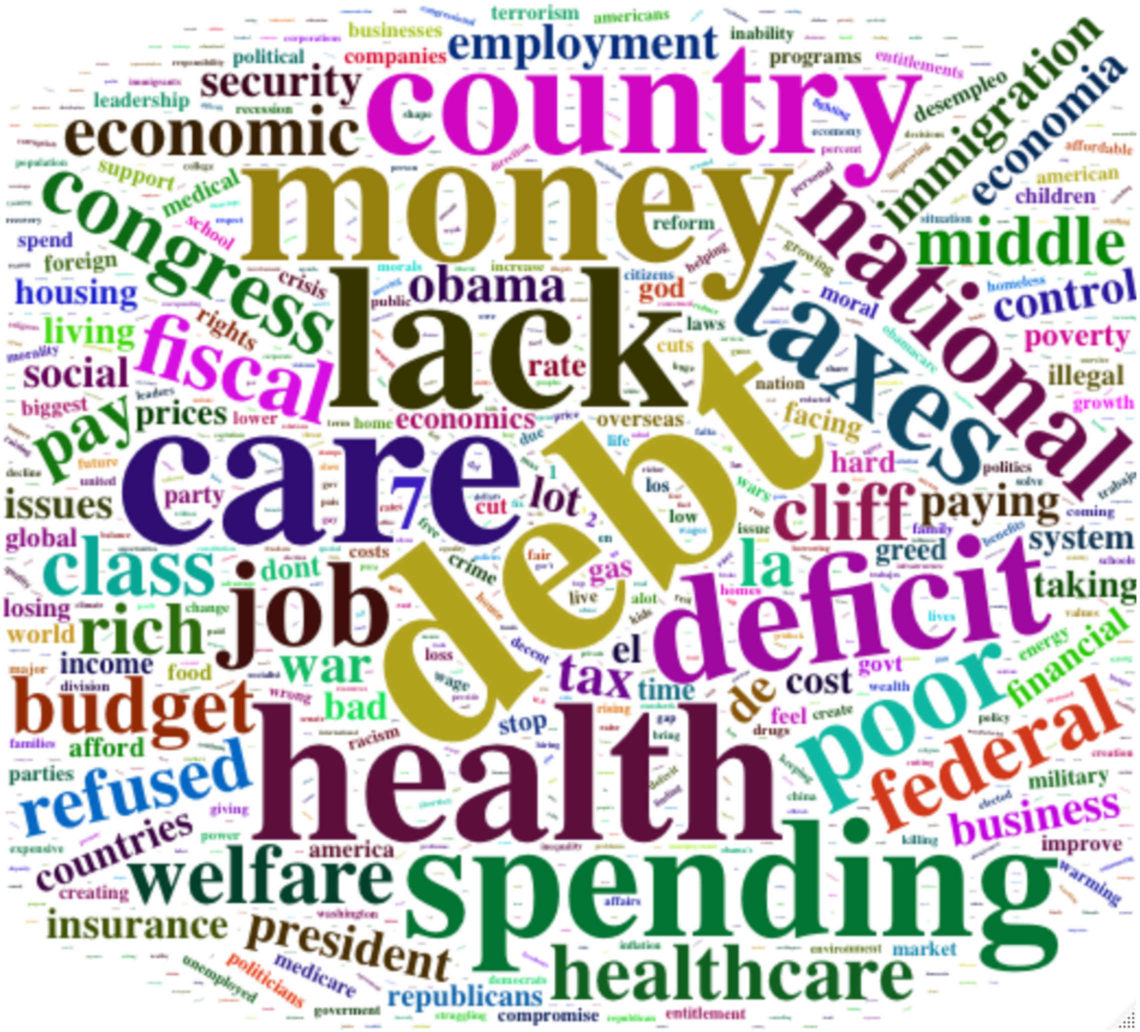

Figure 4.1: A word cloud of the top words contained within the ANES data. The size of each word represents the frequency of appearance.

One public opinion question involved in the survey considers whether respondents approve or dissaprove of the way president Barack Obama was handling the job of president at the time. For this simulation, we estimate the proportion of the population within each state that approve. In other words, we consider a binary response. We treat the original ANES sample as our population and take a subsample with probability to proportional size sampling using the Poisson method (Brewer et al., 1984) with an expected sample size of 1,000. For our size variable, we use the original survey weight plus 0.7 if the true response is "approve" in order to explicitly generate an informative sample. We fit the BUDIS model using $\boldsymbol{\psi}_i = (I(w_{i1}), \ldots, I(w_{i1000}))$ as the input into the ELM, where $I(w_{ij})$ indicates whether or not the jth most frequently occurring word appeared in the free response of unit i. For the linear component, $\boldsymbol{x}_i$, we use indicators of Hispanic ethnicity and gender as poststratification variables, as well as a set of spatial basis functions. We use the first 25 eigenvectors of the state adjacency matrix as our basis functions, although other basis functions could be substituted here. We generate our hidden weights in the matrix $\boldsymbol{A}$ from the standard Normal distribution, and then randomly set 10% equal to zero and we use a vague prior distribution by setting $a = b = 0.5$ and $\sigma^2_\beta = 1000$. We have found that in general the model is not overly sensitive to the choice of distribution for the random weights (i.e. the generating distribution for the matrix $\boldsymbol{A}$), however, depending on the application, it may be desirable to select the distribution through cross-validation. Lastly, we set the number of hidden nodes $h = 240$. This number was chosen such that further increases resulted in little discernible difference in prediction. We also compare to a model that does not use the text data or the ELM component, which we denote Pseudo-likelihood logistic regression (PLLR), as well as both weighted and unweighted direct estimators. The two model based approaches both use post-stratification by sampling from the posterior distribution of the parameters, and then generating estimates for the response value of all units in the population. We repeat the sampling and model fitting procedure 50 times.

Table 4.1 shows the MSE and squared bias of each of the estimators through the

Table 4.1: Simulation results: MSE and squared bias of the four estimators

Estimator	MSE	Bias^2
BUDIS	$\mathbf{1.99 \times 10^{-2}}$	6.43×10^{-3}
PLLR	2.21×10^{-2}	8.09×10^{-3}
Direct	4.49×10^{-2}	$\mathbf{4.72 \times 10^{-3}}$
UW Direct	4.21×10^{-2}	1.33×10^{-2}

simulation. The choice of MSE as our metric in this simulation was based on the goal of SAE. In other cases, interest may lie in inference or individual prediction, in which case a metric that evaluates individual predictions rather than aggregations may be preferable. The first thing to note is that the much higher bias of the unweighted (UW) direct estimator when compared to the direct estimator indicates that the sampling was indeed informative. The two model based approaches were able to handle the informative sampling mechanism through the use of the pseudo-likelihood and improve the MSE dramatically compared to the direct estimates. The BUDIS model was able to further improve upon the PLLR model by reducing MSE about 10% and reducing squared bias around 21%. It is clear in this case that the inclusion of nonlinear modeling of the text covariates results in better estimates. There are other potential text-based models that could be worth comparing to. For example, one could envision a naive model where the vector $\boldsymbol{\psi}_i$ is plugged directly into the model without the use of the ELM. This would result in an extremely large parameter space relative to the sample size considered here, thus variable selection would be of utmost importance. In contrast, for this example, the ELM provides implicit dimension reduction.

4.4 ANES Data Analysis

In order to illustrate this methodology on a real application, we use the entire 2012 ANES dataset to create estimates under the BUDIS model. The total sample size for this dataset was 5,878, with state sample sizes ranging from 4 (Wyoming) to 742 (California). Similar to the simulation study considered in Section 4.3, we estimate the proportion of voting age residents within each state that approved of Barack Obama's job as president at the time the survey was taken. The covariates and hyperparameters were also the same as those considered in the simulation study.

We compare the estimates under the BUDIS model to the direct estimates in Figure 4.2. Note that many of the direct estimates fall towards the extremes due to limited sample sizes in some states. In contrast to this, the model based estimates fall in a narrower range due the effect of "borrowing information" across states. For the most part the spatial pattern under the BUDIS model is as expected. The more traditionally conservative states in the South and towards the Dakotas tend to have lower estimates of approval than the coastal parts of the country. The Northwest portion of the country has a couple unexpected estimates, namely Wyoming and Washington. The higher than expected estimate for Wyoming is likely due to the limited sample size pulling the estimate upward towards the national average, although an effort to find more suitable spatial basis functions could also aid improvement in this area.

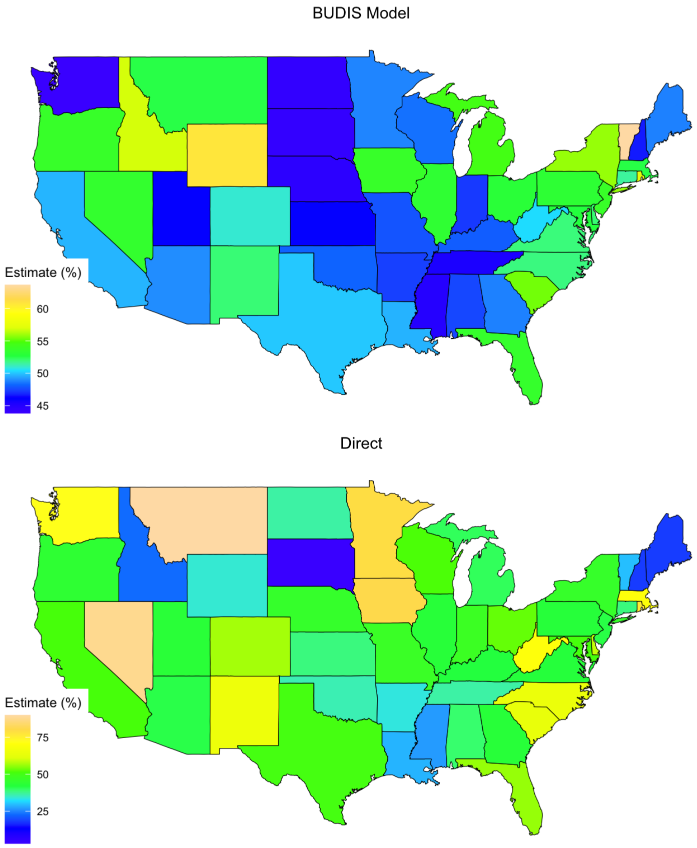

Figure 4.2: Comparison of BUDIS model and direct estimates on 2012 ANES data. Note that separate scales are used in order to emphasize the spatial pattern under each approach.

This example emphasizes how the BUDIS model can be used to construct better estimates of public opinion through the use of complex data types such as free response text. The ANES dataset has a sufficient sample size to construct direct estimates at the national level, but many of the state sample sizes are extremely small, leading to very poor estimates. In this specific case, many of the state level direct estimates fall far away from from the national average. In contrast, the BUDIS model is able to smooth many of these extreme estimates by relying on a model that accounts for spatial dependence between survey respondents as well as the complex free response covariate information. Despite the small state sample sizes, this model is able to yield much more reasonable estimates, such as the general trend of lower approval in the South and higher approval on the East coast. These types of estimates could serve useful for targeting of campaign funds. For instance, in the key states of Florida and Michigan, the model based estimates indicate that Michigan may be more competitive. Furthermore, this example only considers estimates for a single public opinion question, but the ANES survey contains many more public opinion questions that may be of interest to others.

4.5 Discussion

In order to use complex unit level survey data as a covariate for small area estimation, we develop a couple important innovations to the PL unit-level model. The first innovation is the use of deep learning to model nonlinear functions of the complex covariates. This is achieved through the use of random weight methodologies, specifically the ELM. By taking this approach we are able to side-step the need for gradient descent techniques that are typically used in deep learning. In addition, this approach is highly computationally efficient, as it is linear in the parameters that are estimated. Further efficiency is gained through the use of a variational Bayes model fitting procedure.

The second innovation is the use of an imputation model that assigns sample covariates to population units while adjusting for the sampling design. Although this approach is relatively straightforward, modeling the population covariates explicitly could be very burdensome for high-dimensional data and this approach provides a path forward. The use of more advanced approaches to this imputation problem is subject to future work.

In addition to the novel modeling approaches explored here, this work highlights the need to collect more complex data types within surveys. Typical surveys include relatively simple data types such as binary and categorical measurements. However this work shows that more complex data types such as text or functional data may be used to improve the precision of survey based estimates. Currently, federal agencies spend significant resources converting open responses into simple categorical variables. Through the use of our proposed model, or extensions thereof, agencies may be able to rely less on these resources while simultaneously extracting more information from the raw data. The ANES data considered in our examples was chosen in part because of its public availability, in order to limit the need for disclosure issues. However, our approach could be immediately (or with minor modifications) applicable to other complex survey datasets, such as NHANES physical activity monitor data, or web-based respondent tracking.

Chapter 5

A Bayesian Functional Data Model for Surveys Collected under Informative Sampling with Application to Mortality Estimation using NHANES

5.1 Introduction

The use of functional data as either a response or covariate has seen wide usage in recent years. Applications that utilize functional data include longitudinal data analysis (Yao et al., 2005), ecology (Yang et al., 2013), small area estimation (Porter et al., 2014), as well as many others. However, typical models for functional data typically assume a sample that is representative of the population, and thus are not directly applicable to many survey datasets, especially under informative sampling.

For example, the National Health and Nutrition Examination Survey (NHANES) contains functional data in the form of activity monitor curves, yet the survey design is complex leading to a sample that is not representative of the population.

There is a breadth of literature around functional data analysis (FDA). A seminal work in FDA is that of Ramsay and Silverman (2005), while Hsing and Eubank (2015) and Kokoszka and Reimherr (2017) provide more recent treatments of the topic. For further information on the general field of FDA, see the reviews by Morris (2015) and Wang et al. (2016) along with the references therein. The literature on functional data analysis for survey data is quite sparse. Savitsky and Toth (2016) consider the case of functional responses under informative sampling. They treat the response as a Gaussian process to handle functional dependence while simultaneously using a weighted Bayesian pseudo-likelihood to account for informative sampling. One drawback of this approach is that computation under the Gaussian process formulation can become prohibitively difficult in high-dimensional settings.

More recently, Leroux et al. (2019) explore scalar on function regression to predict 5-year mortality rate based on NHANES physical activity covariates. Their approach employs existing software packages that use the survey weights to construct appropriate point estimates but are unable to give correct estimates of uncertainty based on the sample design. The authors state that a resampling procedure may be used to give appropriate standard errors, but the approach is beyond the scope of their work. Ultimately, their use of scalar on function regression was more exploratory and not intended to fully account for the survey design.

In this work, we develop a Bayesian model for scalar on function regression of survey data under informative sampling. Through the use of a Bayesian pseudo-likelihood (Savitsky and Toth, 2016), we are able to give appropriate measures of uncertainty. In addition, we use data augmentation to ensure that the model can be fit in an efficient manner via Gibbs sampling. We also provide an extension to Multinomial response data, which allows for certain multivariate problems to fit into our framework. Similar to Leroux et al. (2019), we are primarily motivated by the topic of mortality estimation with NHANES physical activity covariates, though we note that this methodology is generally applicable to any type of functional survey data. The remainder of this work is outlined as follows. In Section 5.2 we describe our methodology along with necessary background material. Section 5.3 outlines the motivating NHANES dataset. In Section 5.4 we conduct a synthetic data simulation as well as an empirical simulation study that utilizes the public-use NHANES activity monitor data. We also present a data analysis of the public-use NHANES data in Section 5.5. Finally, we provide concluding remarks and discussion in Section 5.6.

5.2 Methodology

5.2.1 Informative Sampling

In many survey data settings, there is dependence between a unit's probability of selection and the response of interest. This is termed *informative sampling* and is known to introduce bias into the model when ignored. Thus, in survey data settings, it is important to account for the survey design in some manner in order to eliminate or reduce this bias. In other words, complex sample designs can lead to samples that are unrepresenative of the population and, thus, the sample model should be adjusted in some way to account for this.

Chapter 1 gives an overview of various methods to account for informative sampling. Of primary interest is the pseudo-likelihood (PL) method introduced by Skinner (1989) and Binder (1983). This approach adjusts the likelihood function by exponentially weighting each unit's likelihood contribution by the corresponding survey weight (i.e. the inverse of the selection probability),

$$\prod_{i \in \mathcal{S}} f(y_i \mid \boldsymbol{\theta})^{w_i}, \tag{5.1}$$

where $\mathcal{S}$ indicates the sample, y_i represents the response value for unit i with survey weight w_i. In a frequentist setting, this PL can be maximized to give a point estimate for $\boldsymbol{\theta}$, however more complex procedures are necessary to give appropriate estimates of uncertainty.

Savitsky and Toth (2016) show that a PL may also be used in a Bayesian framework. In particular, they show that under informative sampling, the use of a PL along with a prior specification leads to a pseduo-posterior distribution,

$$\hat{\pi}(\boldsymbol{\theta}|\mathbf{y}, \tilde{\mathbf{w}}) \propto \left\{ \prod_{i \in \mathcal{S}} f(y_i|\boldsymbol{\theta})^{\tilde{w}_i} \right\} \pi(\boldsymbol{\theta}),$$

that converges to the population posterior distribution. In this scenario, it is important to scale the weights to sum to the sample size in order to attain the appropriate estimates of uncertainty. These scaled weights are represented by $\tilde{w}_i$. This formulation applies generally to Bayesian models and is the approach we use herein to account for informative sampling.

5.2.2 Non-Gaussian Data

Modeling non-Gaussian data types in a Bayesian setting can be computationally burdensome, especially while accounting for informative sampling. Chapter 3 utilizes a data augmentation approach to construct a flexible mixed model for Binomial and Multinomial data under informative sampling. The model for Binomial data is given by,

$$
\begin{aligned}
\boldsymbol{Z}|\boldsymbol{\beta},\boldsymbol{\eta} &\propto \prod_{i \in S} \text{Bin}\left(Z_i|n_i,p_i\right)^{\tilde{w}_i} \\
\text{logit}(p_i) &= \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{\phi}_i'\boldsymbol{\eta} \\
\boldsymbol{\eta}|\sigma_\eta^2 &\sim \text{N}_r(\boldsymbol{0_r}, \sigma_\eta^2 \boldsymbol{I}_r) \\
\boldsymbol{\beta} &\sim \text{N}_q(\boldsymbol{0_q}, \sigma_\beta^2 \boldsymbol{I}_q) \\
\sigma_\eta^2 &\sim \text{IG}(a,b) \\
\sigma_\beta, a, b &> 0,
\end{aligned}
\tag{5.2}
$$

where Z_i represents the response value for unit i in the sample. In this case, $\boldsymbol{x}_i$ is a vector of fixed effects covariates and $\boldsymbol{\phi}_i$ represents a set of spatial basis functions.

In order to fit this model in a computationally efficient manner, Pólya-Gamma data augmentation is used. Specifically, letting $\text{PG}(\cdot,\cdot)$ represent a Pólya-Gamma distribution, Polson et al. (2013) show that

$$
\frac{(e^\psi)^a}{(1+e^\psi)^b} = 2^{-b} e^{\kappa\psi} \int_0^\infty e^{-\omega\psi^2/2} p(\omega) d\omega,
$$

where $\kappa = a - b/2$ and $p(\omega)$ is a $\text{PG}(b,0)$ density. They further show that $(\omega|\psi) \sim$

PG(b, ψ). The PL in (5.2) can be written,

$$\prod_{i \in \mathcal{S}} \left(\frac{(e^{\psi_i})^{Z_i}}{(1 + e^{\psi_i})^{n_i}} \right)^{\tilde{w}_i} = \prod_{i \in \mathcal{S}} \frac{(e^{\psi_i})^{Z_i^*}}{(1 + e^{\psi_i})^{n_i^*}},$$

where $\psi_i = \text{logit}(p_i)$, $Z_i^* = Z_i \times \tilde{w}_i$, and $n_i^* = n_i \times \tilde{w}_i$. This allows for data augmentation of a latent Pólya-Gamma random variable that leads to conjugate Normal priors on the regression parameters.

The Binomial model in (5.2) can also be extended to Multinomial or Categorical data. Following Linderman et al. (2015), the Multinomial distribution with C categories can be rewritten as

$$\text{Multinomial}(\boldsymbol{Z} | n, \boldsymbol{p}) = \prod_{c=1}^{C-1} \text{Bin}(Z_c | n_c, \tilde{p}_c),$$

where

$$n_c = n - \sum_{j<c} Z_j, \quad \tilde{p}_c = \frac{p_c}{1 - \sum_{j<c} p_j}, \quad c = 2, \ldots, C.$$

In this light, a series of $C - 1$ Binomial models may be fit to estimate the parameters for a Multinomial data model.

The modeling framework of Chapter 3 is useful for fitting Binomial data under informative sampling, such as the NHANES mortality data of interest; however, the approach must be extended in order to consider functional covariates.

5.2.3 Functional Covariates

Consider the case where we have J functional covariates and $\kappa_{ij}(t), t \in \mathcal{T}$ denotes the jth functional covariate ($j = 1, \ldots, J$) for unit i at time t. In our case, the domain is time, though other domains may be appropriate depending on the type of functional data. Then, (5.2) can be extended for functional covariates by letting

$$\text{logit}(p_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + \sum_{j=1}^{J} \int_{\mathcal{T}} \eta_j(t)\kappa_{ij}(t)dt,$$

where $\eta_j(t)$ is a functional regression parameter associated with functional covariate j. In what follows, we will assume $J = 1$ (and drop the subscript j), as is the case in our example, but we note that the approach is still applicable for $J > 1$.

In order to reduce the dimension of the problem, we can use a basis expansion representation. In particular, let $\{\phi_k(t) : k = 1, 2, \ldots\}$ be a complete orthonormal basis of the domain $\mathcal{T}$. Then, we can represent the functional covariate as

$$\kappa_i(t) = \sum_{k=1}^{\infty} \xi_i(k)\phi_k(t)$$

and

$$\eta(t) = \sum_{k=1}^{\infty} b(k)\phi_k(t),$$

where $\xi_i(k)$ and $b(k)$ are the expansion coefficients for $\kappa_i(\cdot)$ and $\eta(\cdot)$ respectively. Now, appealing to orthonormality,

$$\text{logit}(p_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + \int_{\mathcal{T}} \eta(t)\kappa_i(t)dt = \boldsymbol{x}_i'\boldsymbol{\beta} + \sum_{k=1}^{\infty} b(k)\xi_i(k). \tag{5.3}$$

Note that any orthonormal basis may be used here, though we explore the use of functional principal components selected through the fast covariance estimation (FAST) approach (Xiao et al., 2016). This is easily implemented via the use of the `refund` package in `R` (Goldsmith et al., 2019).

In practice, the summation in (5.3) is truncated to K. For our purposes, we truncate the summation (i.e., choose K) such that the retained components explain 95% of the variation in the functional data. The choice of the number of principal components to include is not trivial, with many different guidelines available. Our general strategy it to select a number that is "large enough" and to shrink any coefficients corresponding to noise, rather than to eliminate the noise components before inclusion in the model. This results in a finite, though potentially large, number of basis functions. Furthermore, any given basis function may not necessarily be related to the response. Thus, we require some form of variable selection and shrinkage estimator. By doing so, the variable selection prior is able to determine which components of the variation in functional data are correlated with the response. This is similar to the approach taken by Holan et al. (2010).

5.2.4 Horseshoe Prior

In order to provide shrinkage to our functional regression coefficients, we utilize the Horseshoe prior introduced by Carvalho et al. (2010). Although many other methods of Bayesian variable selection exist, this has the advantage of being fully specified, without requiring hyperparameter selection, as well as providing minimial shrinkage to strong signals while still providing a high degree of shrinkage for noise. To implement the Horseshoe prior for (5.3), we use the following hierarchy,

$$\begin{aligned} b(k)|\lambda_k, \tau &\overset{ind}{\sim} \mathrm{N}(0, \lambda_k^2 \tau^2),\ k = 1, \dots, K \\ \lambda_k &\overset{ind}{\sim} \mathrm{C}^+(0,1) \\ \tau &\sim \mathrm{C}^+(0,1), \end{aligned}$$

where $\mathrm{C}^+(\cdot,\cdot)$ represents the Cauchy density truncated below at zero. This prior is considered a global-local shrinkage approach. This can be seen by recognizing that τ applies to all regression parameters and determines the overall level of shrinkage, whereas λ_k is local and applies to a specific coefficient. In this way, coefficients corresponding to noise can attain a higher degree of shrinkage than those with strong signals.

The half-Cauchy priors used in the Horseshoe are not conjugate. However, Makalic and Schmidt (2015) use a data augmentation approach to allow for Gibbs sampling within the Horshoe prior framework. In particular,they use a scale mixture representation of the half-Cauchy such that when $x \sim \mathrm{C}^+(0, A)$, then $x^2|a \sim \mathrm{IG}(1/2, 1/a)$ and $a \sim \mathrm{IG}(1/2, 1/A^2)$, where $\mathrm{IG}(a, b)$ represents the Inverse Gamma distribution with shape parameter a and scale parameter b. This leads to an alternate formulation of the Horseshoe prior hierarchy,

$$\begin{aligned} b(k)|\lambda_k, \tau &\overset{ind}{\sim} \mathrm{N}(0, \lambda_k^2 \tau^2),\ k = 1, \dots, K \\ \lambda_k^2|\nu_k &\overset{ind}{\sim} \mathrm{IG}(1/2, 1/\nu_k) \\ \tau^2|\nu_\tau &\sim \mathrm{IG}(1/2, 1/\nu_\tau) \\ \nu_1, \dots, \nu_K, \nu_\tau &\overset{ind}{\sim} \mathrm{IG}(1/2, 1), \end{aligned}$$

that allows for straightforward Gibbs sampling.

5.2.5 Functional Data Model under Informative Sampling

We now present our model for non-Gaussian data under informative sampling with functional covariates, which makes use of the modeling elements discussed so far:

$$\begin{aligned}
\boldsymbol{Z}|\boldsymbol{\beta},\boldsymbol{\eta} &\propto \prod_{i\in S} \mathrm{Bin}\left(Z_i|n_i,p_i\right)^{\widetilde{w}_i} \\
\mathrm{logit}(p_i) &= \boldsymbol{x}_i'\boldsymbol{\beta} + \sum_{k=1}^{K} b(k)\xi_i(k) \\
\boldsymbol{\beta} &\sim \mathrm{N}_q(\boldsymbol{0_q},\sigma_\beta^2\boldsymbol{I}_q) \\
b(k)|\lambda_k,\tau &\overset{ind}{\sim} \mathrm{N}(0,\lambda_k^2\tau^2),\ k=1,\ldots,K \qquad (5.4) \\
\lambda_k^2|\nu_k &\overset{ind}{\sim} \mathrm{IG}(1/2,1/\nu_k) \\
\tau^2|\nu_\tau &\sim \mathrm{IG}(1/2,1/\nu_\tau) \\
\nu_1,\ldots,\nu_K,\nu_\tau &\overset{ind}{\sim} \mathrm{IG}(1/2,1) \\
\sigma_\beta^2 &> 0.
\end{aligned}$$

In this model, $\boldsymbol{x}_i$ represents a q-dimensional vector of scalar covariates and $\xi_i(k)$ represents the kth basis expansion coefficient for observation i.

This model makes use of a Bayesian pseudo-likelihood to account for informative sampling, allowing for population level inference. We also make use of Pólya-Gamma data augmentation for efficient Gibbs sampling. If desired, prior information on the scalar covariates may be incorporated through the selection of σ_β^2, though we use a relatively diffuse prior by letting $\sigma_\beta^2 = 10$. The full conditional distributions are given in Appendix A.3.

It is straightforward to implement the model in the Multinomial data setting through the use of a stick-breaking representation, as discussed in Section 5.2.2. It would also be straightforward to use a Gaussian pseudo-likelihood in place of the Binomial one given here, as conjugacy would be retained.

5.3 NHANES Data Description

The NHANES is a survey conducted by the National Center for Health Statistics that utilizes a complex survey design to collect health and nutrition data in the United States. Of primary interest to us is the physical activity monitor (PAM) data collected during the 2003-2004 and 2005-2006 samples. Along with this, we are interested in mortality as a response value.

NHANES provides microdata to the public, however, a substantial amount of data processing is required to utilize the data for inference. Leroux et al. (2019) provide very helpful exposition on processing the data as well as the `rnhanesdata` package in `R` for doing so. All analyses in this work were conducted using data that was prepared and processed in the same manner as Leroux et al. (2019).

In particular, we use the NHANES samples from 2003-2004 and 2005-2006, as these contain the PAM data of interest. For each minute during a seven consecutive day period, the data contains an activity intensity value for each subject. Not all subjects were in compliance, resulting in variation in wear-time between subjects. Thus, all subjects that had less than 3 days of 10 hours or greater wear-time were dropped. In addition, subjects outside of the age range 50-85, or who were missing mortality or age data were also dropped. The resulting sample size was 3,208. In addition to PAM data as a functional covariate, we use age as a scalar covariate.

The PAM activity measurements were transformed using $f(x) = \log(1 + x)$. The subjects in the sample had a varying numbers of days with activity data. To account for this, we use the PAM data averaged across days within subjects, resulting in a single 24 hour curve for each subject.

The NHANES sample contains the required survey weights. Following Leroux et al. (2019), these are reweighted to account for missing data. More in depth reweighting schemes may be desired in practice. However, discussion on reweighting procedures for missing data is beyond the scope of this work. Thus, for illustration, we do not consider this problem further.

5.4 Simulation Studies

This section includes two different simulation studies. The first is a synthetic data simulation, where we develop a synthetic population based on pre-specified and known functional coefficients. We divide the population into clusters which are used to take a two-stage sample under an informative design. The primary goal of the first simulation is to evaluate the quality of our uncertainty estimates around the functional coefficient.

Our second simulation is an empirical simulation study where we treat the existing NHANES sample as a population and take a further informative subsample. Simulating in this manner allows us to retain many of the characteristics of the NHANES data, while having a known truth to compare out of sample predictions to. Note that in this case, unlike the synthetic data simulation, the true functional coefficient is unknown.

5.4.1 Synthetic Data Simulation

In order to assess the ability of our model to perform inference on a population level functional coefficient, we create a synthetic population constructed under a known functional coefficient. In this case, we consider both a very smooth function (function A) as well as a more variable function (function B). These are shown as dashed lines in Figure 5.1.

We create a population of 100,000 individuals, and assign each individual into one of 100 different clusters. The cluster probabilities are generated from a Dirichlet distribution with all concentration parameters set to one. This results in a population with uneven cluster memberships. To generate the individual functional covariates we consider weighted combinations of the functional covariates observed in the NHANES sample. For each individual, we generate a set of weights by drawing from a Dirichlet distribution with concentration parameters set to 1/3208 (or one over the number of observations in the NHANES data). This results in synthetic curves that place a large amount of weight on a few curves, and little weight on the remaining curves. This process results in curves that look quite similar to the ones observed in the NHANES data, without any two curves being identical. In addition to the functional covariate, we generate one categorical covariate with four categories and corresponding coefficients drawn from a standard normal distribution. Finally, using the synthetic covariates and regression coefficients, we generate a linear predictor for each individual. We apply the inverse logit transformation to the linear predictors in order to generate a mortality probability for each individual, and then generate responses by drawing Bernoulli random variables conditional on the probabilities.

After generating a synthetic population, we are able to take an informative sample that can be used for estimation. In this case, we take a two stage sample. In the first stage, we sample ten of the available clusters. Clusters are sampled with probability proportional to the number of individuals within the cluster. Thus, clusters with more membership are more likely to be included in the sample. Cluster inclusion probabilities may be denoted as p_c. In the second stage, for each cluster chosen in stage one, we use Poisson sampling (Brewer et al., 1984) to retain an expected sample size of 10% of the cluster size. The probability of selection of unit i in cluster c, conditional on cluster c being in the first stage sample, is denoted as $p_{i|c}$. We set this probability to be proportional to $\exp\left\{2 \times I(Z_i = 1) + 2 \times I(X_{i,cat} = 1) + 0.5 \times I(X_{i,cat} = 2)\right\}$, where $I(Z_i = 1)$ is an indicator function indicating that the response for individual i is 1. Similarly, $I(X_{i,cat} = k)$ indicates that the categorical covariate for unit i is equal to k. The use of only categorical variables in the second stage makes this stage similar to a stratified random sample. Inclusion of the response in this second stage forces an informative design. In practice, informativeness is often an unintended consequence of a given sample design, rather than a feature. Finally, after sampling both stages, the marginal inclusion probabilities can be denoted as $p_i = p_{i|c} \times p_c$.

Using this two stage sample design, we take 50 samples each from the synthetic populations under functional coefficients A and B. For each sample, we fit a Bayesian pseudo-likelihood model using functional principal components basis functions (BPL-FPCA) as well as cyclic B-spline basis functions of order 30 (BPL-CBS) similar to Leroux et al. (2019). We also compare directly to the model fit by Leroux et al. (2019), which uses cyclic B-spline basis functions of order 30.

For each sample, we are able to compare the functional coefficient estimates, along with their corresponding uncertainty, to the truth. Table 5.1 provides a summary of these results. We present the mean absolute error (MAE), the absolute bias, and the 95% credible/confidence interval coverage rates. Each metric is averaged across time (the domain of the functional coefficient) as well as across the sampled datasets. In general, the BPL-FPCA model seems to result in slightly worse point estimates in terms of MAE. Although the model used by Leroux et al. (2019) results in low MAE, particularly when the true function is smooth, it severely underestimates the uncertainty in both the smooth and unsmooth settings. In comparison, both Bayesian pseudo-likelihood models result in coverage rates that are much closer to optimal. In general for this simulation, the BPL-CBS model seems to strike a strong balance between selecting good point estimates for the functional coefficient and quantifying the uncertainty appropriately.

Function	Model	MAE	Abs. Bias	Coverage
A	BPL-FPCA	1.7×10^{-3}	5.2×10^{-4}	96.8%
	BPL-CBS	7.9×10^{-4}	$\mathbf{2.4 \times 10^{-4}}$	**94.3%**
	Leroux	$\mathbf{5.7 \times 10^{-4}}$	2.5×10^{-4}	60.7%
B	BPL-FPCA	9.7×10^{-4}	1.7×10^{-3}	**99.2%**
	BPL-CBS	$\mathbf{6.1 \times 10^{-4}}$	$\mathbf{1.7 \times 10^{-3}}$	90.6%
	Leroux	6.2×10^{-4}	1.8×10^{-3}	44.4%

Table 5.1: Synthetic data simulation results comparing mean absolute error (MAE), absolute bias, and 95% credible/confidence interval coverage rate. All results are averaged pointwise along the functional coefficient and across all simulated datasets.

We also compare the true functional coefficients to the average coefficients across the sampled datasets for each model in Figure 5.1. For true function A, the models that use cyclic B-splines follow the truth quite well. The BPL-FPCA model, while able to capture the general trend, has additional "noise" that increases the MAE.

In setting B, the cyclic B-spline models are able to capture the general trend, but unable to follow some of the peaks and valleys in the true function. In contrast, the BPL-FPCA model is better able to capture some of these high and low points (though not perfectly).

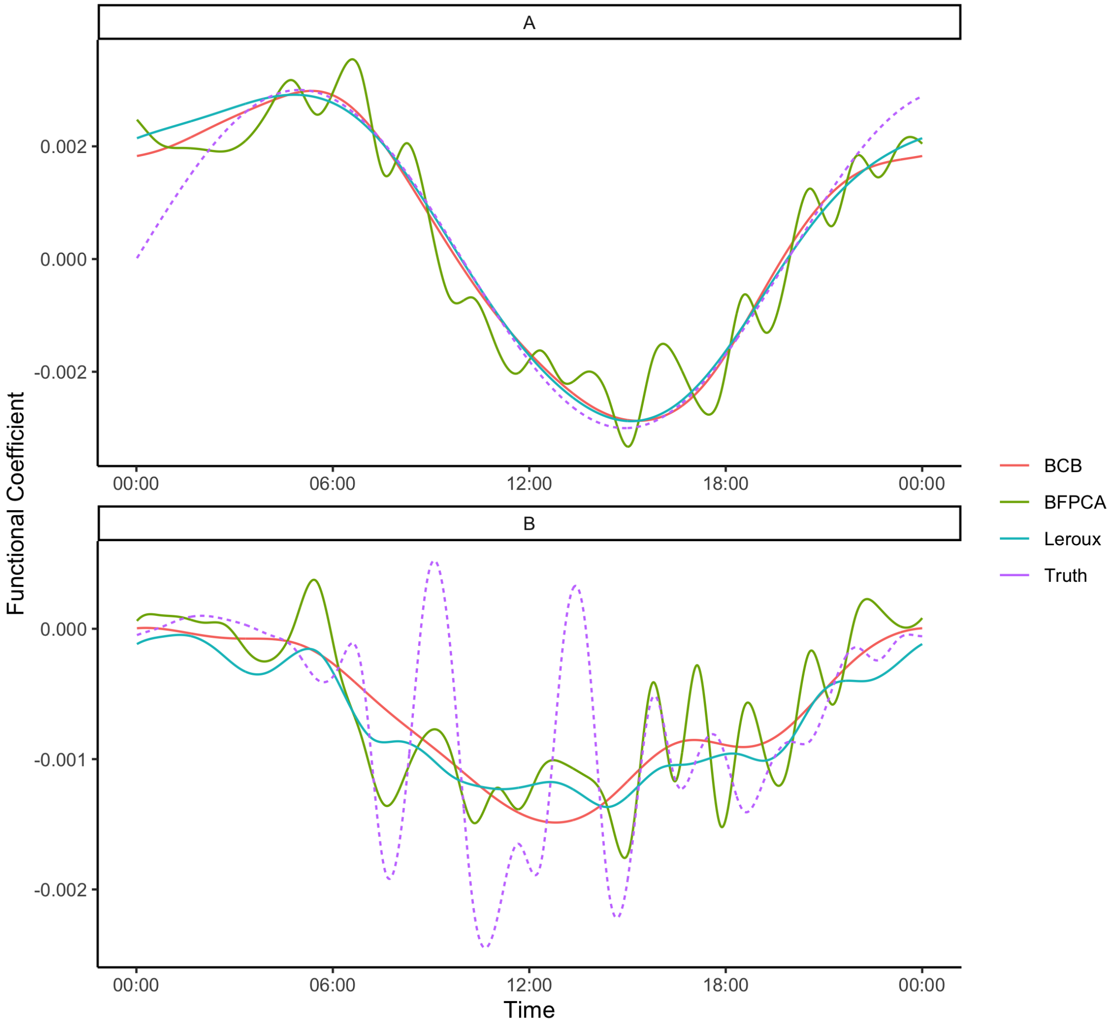

Figure 5.1: Comparison of true functional coefficient to estimated functional coefficient (averaged across simulated datasets). The true curves are shown as dashed lines.

The simulation results do not give indication that there is a clear favorite in terms of which set of basis functions to use. Rather, it most likely depends on the properties of the data at hand as well as the goals of the analysis. Our methodology is not beholden to any given set of basis functions. It is straightforward to swap the FPCA and cyclic B-spline bases for others and we encourage practitioners to explore the effects of these decisions.

5.4.2 Empirical Simulation

The goal of this empirical simulation is twofold. First, we want to confirm that the model is able to adjust for an informative sampling mechanism in order to allow for population level inference. Second, we want to assess whether or not the use of functional covariates, as given in the model, leads to improved estimates for units in the population.

To design such a simulation, we begin by treating the existing NHANES sample data as our population. This provides a baseline truth for which we can compare to. Next, we subsample from the NHANES data in an informative manner. Doing so, we are able to fit the model using the subsampled data and then compare to the population truth (i.e. the original sample data). To take this subsample, we use probability proportional to size sampling via the Poisson method (Brewer et al., 1984) with an expected sample size of 500. We construct the size variable as $s_i = \exp\{w_i^* + 2 * I(Z_i = 1)\}$ where w_i^* is the NHANES reported survey weight after scaling to have mean zero and variance 1, and $I(Z_i = 1)$ is an indicator that the ith respondent died within 5 years of the survey. Through this subsampling procedure, we obtain a new set of weights that are the inverse probabilities of selection.

The response data of interest is the binary indicator of 5-year mortality. We use age along with an intercept term as a scalar covariate and the PAM data as a functional covariate. After subsampling, we fit both the BPL-CBS and BPL-FPCA models as well as the model used by Leroux et al. (2019). We also fit a basic version of the Bayesian pseudo-likelihood model that uses only the scalar covariate and disregards the functional data (BPL-S). In addition we implement unweighted versions of the BPL-FPCA and BPL-S models (UW-FPCA and UW-S respectively).

After fitting each of these models, we are able to make mortality predictions for the entire population. This results in a 5-year probability of mortality for each person in the population that we can compare to their actual mortality result. Because these outcomes are binary, we use binary cross-entropy (BCE) as a loss function to compare rather than mean-squared error. This is calculated as,

$$\text{BCE} = -\frac{1}{N}\sum_{i=1}^{N} Z_i \text{log}(\hat{p}_i) + (1 - Z_i)\text{log}(1 - \hat{p}_i),$$

where $\hat{p}_i$ is the posterior mean probability of mortality for unit $i = 1, \ldots, N$ in the population with size N. Note that a lower value of BCE indicates a better model fit.

We repeat this subsampling and model fitting procedure 50 times, resulting in a distribution of BCE loss values under each model. We compare these distributions in Figure 5.2. It is immediately clear that the two unweighted models perform much worse than the weighted models. These unweighted models do not account for the informative sample design and thus introduce a large amount of bias when making inference on the population. The weighted models are able to account for the sample design, and thus result in much lower values of BCE. Additionally, the distributions of BCE under the BPL-CBS and BPL-FPCA models are shifted to the left of BPL-S, indicating that the functional covariate does aid in prediction of mortality for members of the population. Interestingly, the Leroux model is only slightly better than the BPL-S model, perhaps due to oversmoothing of the functional coefficient. These results indicate that for population level inference based on the full sample data, we should use a model that utilizes the functional covariates while also accounting for the survey design through a Bayesian pseudo-likelihood.

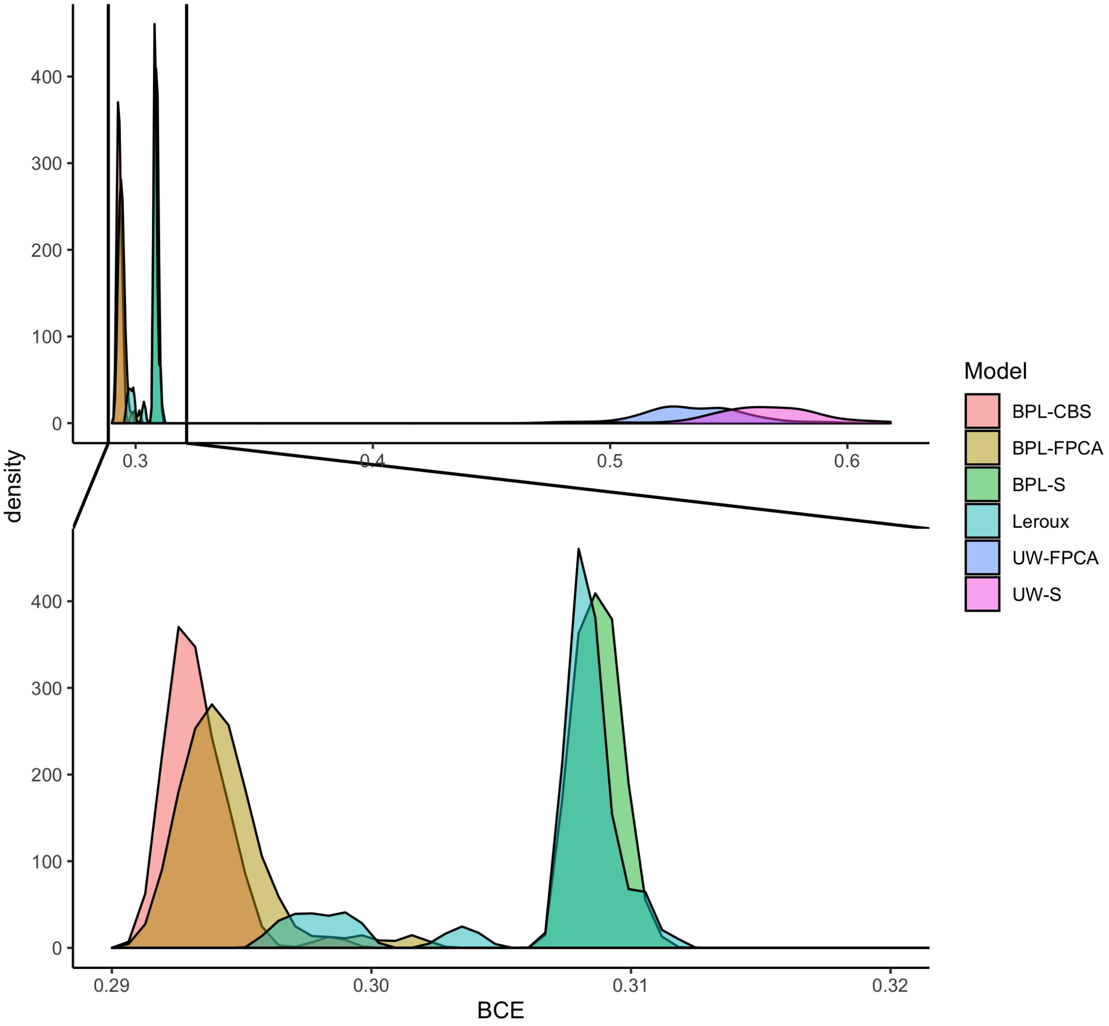

Figure 5.2: Simulation based distribution of BCE values under each model. The lower subplot focuses on the four weighted models which have substantially lower BCE.

5.5 NHANES Data Analysis

5.5.1 5-Year Mortality Estimate

Using our Bayesian pseudo-likelihood based model for functional covariates with cyclic B-splines (BPL-CBS), we now analyze the NHANES PAM data and its relationship with mortality. We use the same dataset considered in the simulation study and outlined in Section 5.3, with a sample size of 3,208. We treat the 5-year mortality indicator as our binary response, and use age, gender, body mass index (BMI), race, education level, as well as self reported presence of a mobility problem, diabetes, coronary heart disease (CHD), congestive heart failure (CHF), cancer, and stroke as our scalar covariates in addition to the PAM data as a functional covariate. Note that these are the same covariates considered by Leroux et al. (2019).

We fit the model via Gibbs sampling with 5,000 iterations and discard the first 1,000 iterations as burn-in. Convergence was assessed via traceplots of the sample chains, where no lack of convergence was detected.

After fitting the model, we are able to make population level inference. We plot the posterior mean of the functional regression coefficient, $\eta(t)$, along with a pointwise 95% credible interval in Figure 5.3. For the most part $\eta(t)$ is estimated to be negative, as expected, indicating that increased levels of activity are associated with lower expected mortality rate. The primary time period where the credible interval does not contain zero is around 12 p.m. The model used by Leroux et al. (2019) underestimates the uncertainty, and correspondingly results in a much larger significant window than the one obtained here.

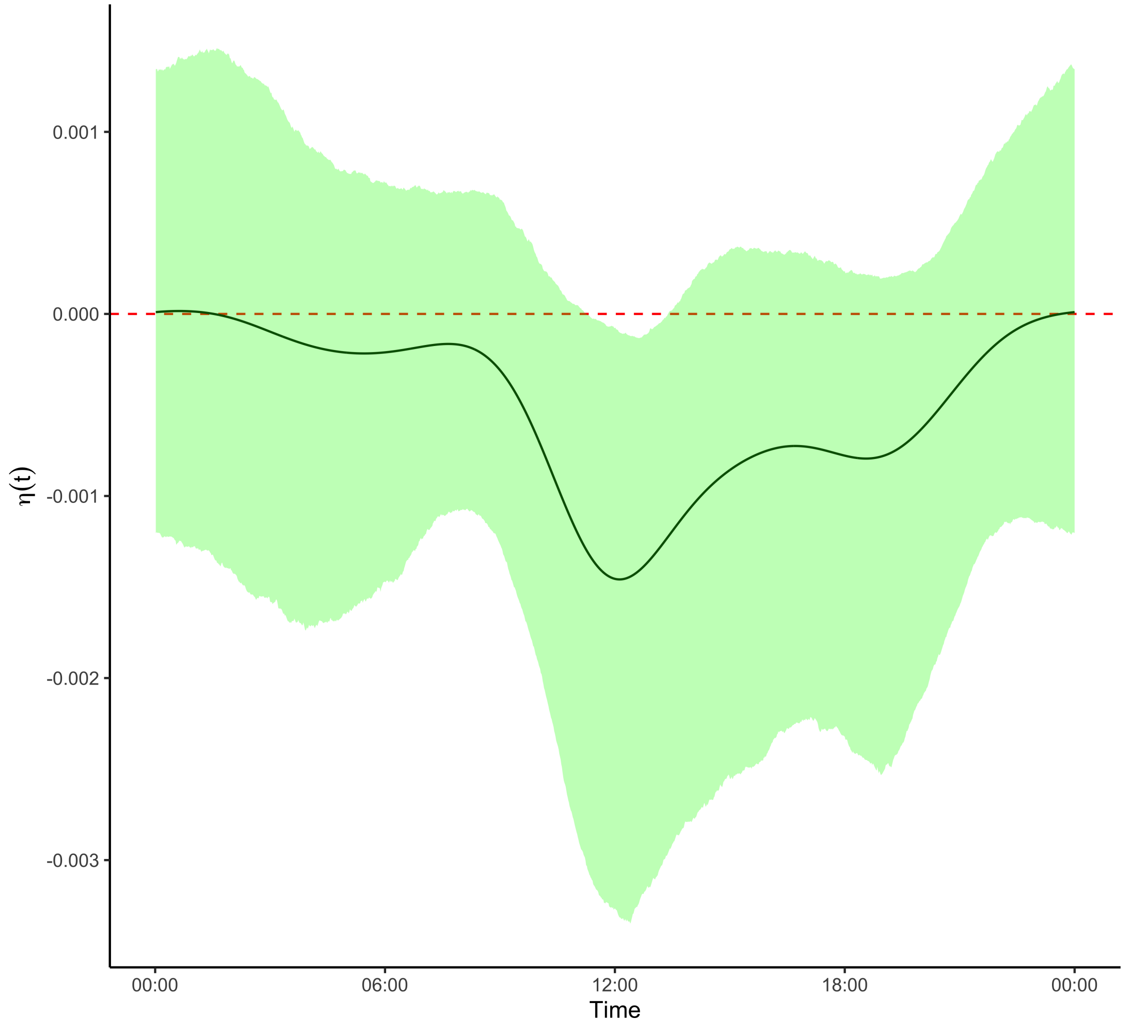

Figure 5.3: Estimate of the PAM functional regression coefficient for 5-year mortality along with pointwise 95% credible interval.

In addition to examination of the functional regression coefficient, we can also glean insight by examining how variation in activity level of individuals changes mortality estimates. In Figure 5.4, we plot activity curves for 3 individuals contained in the NHANES sample along with the accompanying posterior distribution of 5-year mortality rate. Individual A has a very low level of activity, resulting in a high expected mortality rate. However, there is also a great deal of uncertainty around this rate. Individuals B and C both have increasing level of overall activity, especially in the early morning and late afternoon, resulting in decreasing expected mortality rate. As the activity level increases, the uncertainty around the mortality rate tends to decrease.

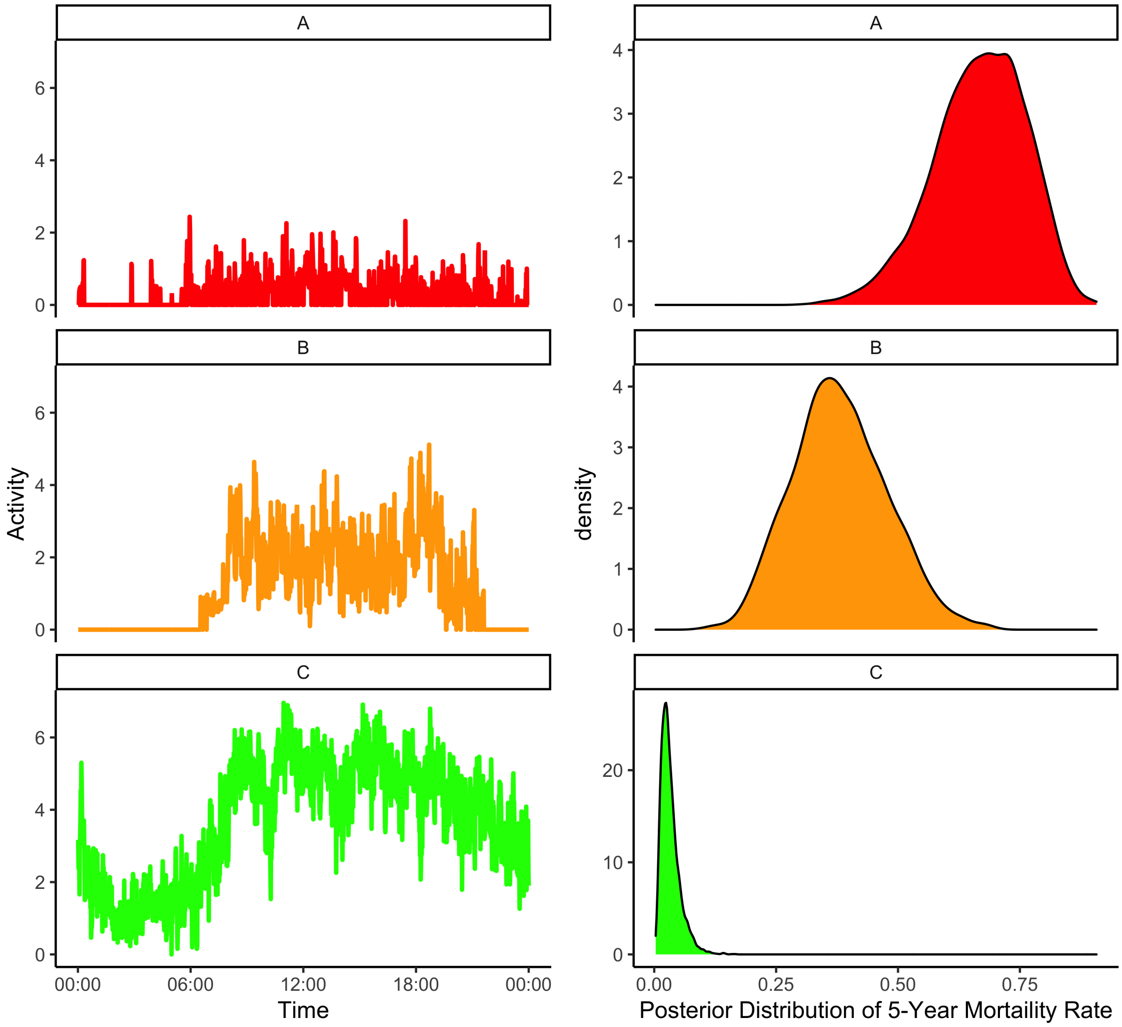

Figure 5.4: Activity curves for 3 individuals contained in the NHANES sample along with posterior distribution of 5-year mortality rate.

5.5.2 Joint Mortality Estimate

In addition to univariate estimation of 5-year mortality, our model allows for joint estimation at multiple time scales through the Multinomial data model. In this case we wish to make joint mortality estimates for years 1-5. To do so, we begin by assigning survey respondents into distinct categories: those who died within one year of the survey, those who died after 1 year but before 2 years, those who died after 2 years but before 3 years, those who died after 3 years but before 4 years, those who died after 4 years but before 5 years, and finally those that did not die before 5 years. Assigning groups in this way results in a Multinomial or Categorical data distribution with 6 categories. Thus, we are able to use the stick breaking representation of the Multinomial distribution in order to fit $C - 1 = 5$ independent Binomial data models that allow us to make joint estimates of mortality at various time points.

We use the same individuals from our 5-year mortality example to examine the effects of activity level on multi-year mortality. Figure 5.5 plots the activity curves for these individuals alongside their posterior mean mortality rates for year 1-5. We also provide 95% credible intervals. Once again, we see that both expected mortality rate and uncertainty increase as activity level generally decreases. Because these estimates are joint, we can also see that decreased activity is associated with steeper marginal increases in mortality for the near future than for years further away from the survey.

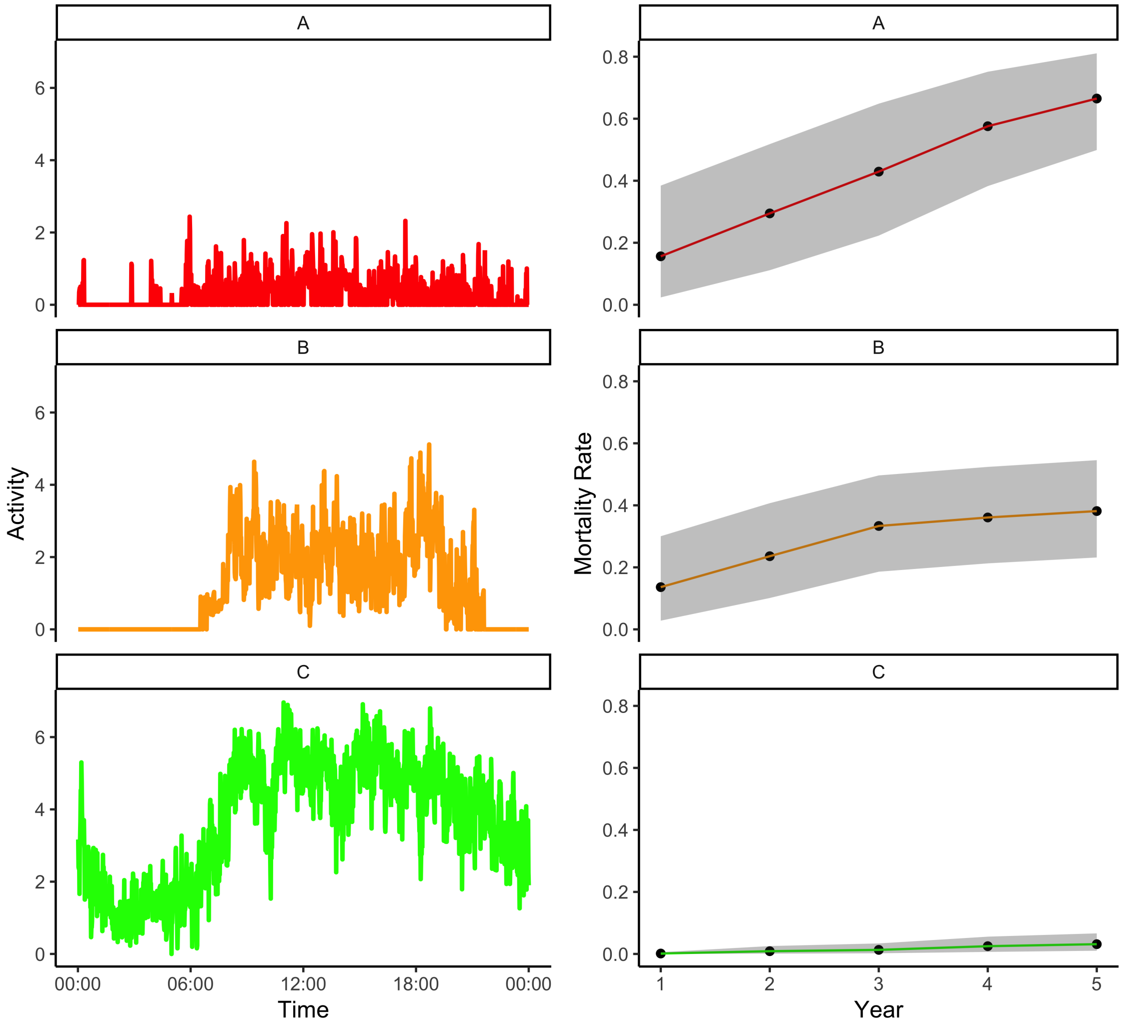

Figure 5.5: Activity curves for 3 individuals along with accompanying posterior mean mortality rate for years 1-5 and 95% credible intervals.

5.6 Discussion

In this work, we develop a Bayesian non-Gaussian data model for functional covariates under informative sampling. We rely on a pseudo-likelihood approach to account for survey design which works in combination with Pólya-Gamma data augmentation to allow for conjugate full conditional distributions of the regression parameters. This method is designed for Binomial or Multinomial data models, though it is straightforward to replace this with a Gaussian data model. Our approach uses an basis expansion representation of the functional covariates alongside the Horseshoe prior to provide regularization. As with the data model, we use a data augmentation approach for the Horseshoe prior, meaning that all full-conditional distributions in the model are conjugate. This allows for straightforward and efficient Gibbs sampling, which can be highly important in high-dimensional settings such as the one explored here.

We conduct both a synthetic data simulation, as well as an empirical simulation study using NHANES data, that show that our approach is able to reduce the bias attributable to informative sampling while also making use of the functional data to improve estimates for members of the population. Importantly, our method is able to give accurate uncertainty quantification as well as provide joint mortality estimates, improving over existing methodology. We also provide a full analysis of the NHANES data that allows us to make inference and prediction on the population. We conduct both a univariate analysis concerning 5-year mortality rate as well as a multivariate analysis concerning years 1-5 mortality rate.

Our methodology extends the literature on functional regression to the survey data setting. The approach is flexible in that users have a choice of data model and basis expansion and also allows for joint estimation of scalar regression coefficients. Currently, there is a limited amount of functional data collected under complex surveys, as analysis options are limited. It is our hope that with the availability of this methodology, collection of functional data via surveys will become more widespread.

Although not explored in this work, similar approaches may be undertaken for function on scalar or function on function regression under complex survey designs. Another potential avenue of future research would involve the use of nonlinear modeling techniques that utilize these same functional covariates. Finally, although we were able to jointly estimate mortality at multiple time points, it would be interesting to explore models that can estimate continuous survival curves based on the NHANES activity data.

Chapter 6

Conclusion

The availability of unit-level approaches to SAE lags that of area-level alternatives. A primary reason for this is the computational constraints that accompany unit-level modeling while considering informative sample designs. This dissertation aims to lessen the gap.

First, a comprehensive review of available mechanisms to account for informative sampling has been provided in Chapter 1. Accounting for sample design is a critical step in the development of any unit-level model in order to reduce or eliminate potential bias. This review gives particular emphasis to the problem of SAE, whereby a subset of methods are compared in an empirical simulation study and application to poverty estimation. These results demonstrate that different approaches to the problem of informative sampling may each yield estimates with reduced MSE compared to direct estimators. However, there are other trade-offs to consider such as computation time, ease of including covariates, and quality of uncertainty estimates.

Chapter 2 has introduced a computationally efficient method to model count data at the unit level. A Bayesian pseudo-likelihood is used to adjust for informative sample designs, and prior distributions are developed based on recent multivariate log-Gamma distribution theory, in order to allow for efficient computation. In addition, an importance sampling step is developed to mitigate issues that arise with zero counts. This methodology is demonstrated through an empirical simulation study for estimation of housing vacancies.

Binary and categorical data are perhaps the most common data types in unit-level survey data. Thus, Chapter 3 has provided methodology for these types of responses, again relying on the Bayesian pseudo-likelihood. An efficient Gibbs sampling scheme is developed for moderately sized data sets, as well as a variational Bayes approximation for extremely large data sets. Along with an empirical simulation study, this chapter illustrates the methodology to the important problem of county level estimation of health insurance rates by IPR category for the entire contiguous United States.

One strength of unit-level modeling is that it can tap into the rich domain of unit-level covariates. Often, complex data structures, such as text and functional data, are collected at the unit level. These data types may lead to higher quality population estimates, or may be useful for inference. Chapter 4 provides an extension of the methodology contained in Chapter 3, that can be useful when complex covariates are used. The extension relies on a type of feedforward neural network known as the extreme learning machine, which allows for nonlinearity and interaction effects. Importantly, this modeling approach fits naturally into the Bayesian pseudo-likelihood framework discussed previously. This methodology is illustrated via text data from the American National Election Studies.

Lastly, Chapter 5 provides methodology for functional covariate data under informative sampling. Using the Bayesian pseudo-likelihood, a basis function expansion is linked with a Bayesian variable selection approach to reduce the dimensionality of the problem and add regularization. The model is illustrated via a simulation and application to mortality estimation using physical activity monitor data from the National Health and Nutrition Examination Survey. Importantly, this work builds on previous work by allowing for the joint estimation of mortality at multiple scales.

The methodology developed in this dissertation relies on the Bayesian pseudo-likelihood to account for the informative sampling design. This is certainly not the only means to doing so, but was chosen for computational reasons. Computation is one of the limiting factors in unit-level modeling, and thus every effort was made to ease this burden. However, there may be cases where a pseudo-likelihood is not appropriate (e.g. a joint model for multiple survey data sets, where parameters are shared between surveys). For this reason, unit-level modeling with other mechanisms to account for sample design is still an important avenue of research.

There are also other avenues of research that may be valuable in expanding the utility of unit-level modeling. This work has exploited spatial modeling (i.e. people in nearby counties may exhibit similar responses), but temporal correlation was not considered, as the relevant applications were generally representing the population at a single point in time. Some survey designs will take repeated measurements on respondents over time, and others may generate new samples at various time points. In these cases, it may be possible to construct temporal or spatio-temporal SAE models at the unit-level that could improve the precision of estimates.

One theme of this dissertation has been that of binary and categorical data. These data types are both frequently occurring in survey data, yet in some cases these data are observed with some error (i.e. misclassified). Although measurement error over continuous survey variables is an active area of research, there is little research devoted to the binary or categorical setting. This is another area where further research could expand the utility of unit-level modeling.

In summary, unit-level modeling of survey data under informative sampling is an important area of research. These models are important to the development of small area estimates, but may also be used for inference. This dissertation has provided several computationally efficient methods for common types of survey data as well as for some non-standard data types. These methods and their potential extensions can help to bridge the gap between the advantages of unit-level modeling in theory and their implementation in practice.

Appendix A

Full Conditional Distributions

A.1 Full Conditional Distributions for Chapter 2

A.1.1 Random Effects

$$\begin{aligned}
\boldsymbol{\eta}|\cdot &\propto \prod_{\ell=1}^{L}\prod_{i\in\mathcal{S}} \exp\left\{\tilde{w}_i z_i^{(\ell)} \boldsymbol{\psi_i'}^{(\ell)} \boldsymbol{\eta} - \tilde{w}_i \exp(\boldsymbol{x_i'}^{(\ell)}\boldsymbol{\beta} + \xi_i^{(\ell)})' \exp(\boldsymbol{\psi_i'}^{(\ell)}\boldsymbol{\eta})\right\} \\
&\quad \times \exp\left\{\alpha \boldsymbol{1_r'} \alpha^{-1/2} \frac{1}{\sigma_k} \boldsymbol{I_r}\boldsymbol{\eta} - \alpha \boldsymbol{1_r'} \exp\left(\alpha^{-1/2}\frac{1}{\sigma_k}\boldsymbol{I_r}\boldsymbol{\eta}\right)\right\} \\
&= \exp\left\{\boldsymbol{\alpha_\eta' H_\eta \eta} - \boldsymbol{\kappa_\eta'} \exp(\boldsymbol{H_\eta \eta})\right\}
\end{aligned}$$

$$\boldsymbol{H_\eta} = \begin{bmatrix} \boldsymbol{\Psi} \\ \alpha^{-1/2}\frac{1}{\sigma_k}\boldsymbol{I_r} \end{bmatrix}, \quad \boldsymbol{\alpha_\eta} = (\boldsymbol{\tilde{w}'} \odot \boldsymbol{Z'}, \alpha \boldsymbol{1_r'})', \quad \boldsymbol{\kappa_\eta} = (\boldsymbol{\tilde{w}'} \odot \exp(\boldsymbol{X\beta} + \boldsymbol{\xi})', \alpha \boldsymbol{1_r'})'$$

$$\boldsymbol{\eta}|\cdot \sim \text{cMLG}(\boldsymbol{H_\eta}, \boldsymbol{\alpha_\eta}, \boldsymbol{\kappa_\eta})$$

$$
\begin{aligned}
\boldsymbol{\xi}|\cdot &\propto \prod_{\ell=1}^{L}\prod_{i\in\mathcal{S}} \exp\left\{\tilde{w}_i z_i^{(\ell)}\xi_i^{(\ell)} - \tilde{w}_i \exp(\boldsymbol{x_i'}^{(\ell)}\boldsymbol{\beta} + \boldsymbol{\psi_i'}^{(\ell)}\boldsymbol{\eta})' \exp(\xi_i^{(\ell)})\right\} \\
&\quad \times \exp\left\{\alpha \boldsymbol{1_n'}\alpha^{-1/2}\frac{1}{\sigma_\xi}\boldsymbol{I_n}\boldsymbol{\xi} - \alpha\boldsymbol{1_n'}\exp\left(\alpha^{-1/2}\frac{1}{\sigma_\xi}\boldsymbol{I_n}\boldsymbol{\xi}\right)\right\} \\
&= \exp\left\{\boldsymbol{\alpha_\xi'}\boldsymbol{H_\xi}\boldsymbol{\xi} - \boldsymbol{\kappa_\xi'}\exp(\boldsymbol{H_\xi}\boldsymbol{\xi})\right\} \\
\boldsymbol{H_\xi} &= \begin{bmatrix} \boldsymbol{I_n} \\ \alpha^{-1/2}\frac{1}{\sigma_\xi}\boldsymbol{I_n} \end{bmatrix}, \quad \boldsymbol{\alpha_\xi} = (\boldsymbol{\tilde{w}'} \odot \boldsymbol{Z'}, \alpha\boldsymbol{1_n'})', \quad \boldsymbol{\kappa_\xi} = (\boldsymbol{\tilde{w}'} \odot \exp(\boldsymbol{X\beta} + \boldsymbol{\Psi\eta})', \alpha\boldsymbol{1_n'})' \\
\boldsymbol{\xi}|\cdot &\sim \text{cMLG}(\boldsymbol{H_\xi}, \boldsymbol{\alpha_\xi}, \boldsymbol{\kappa_\xi})
\end{aligned}
$$

A.1.2 Fixed Effects

$$
\begin{aligned}
\boldsymbol{\beta}|\cdot &\propto \prod_{\ell=1}^{L}\prod_{i\in\mathcal{S}} \exp\left\{\tilde{w}_i z_i^{(\ell)}\boldsymbol{x_i'}^{(\ell)}\boldsymbol{\beta} - \tilde{w}_i \exp(\boldsymbol{\psi_i'}^{(\ell)}\boldsymbol{\eta} + \xi_i^{(\ell)})' \exp(\boldsymbol{x_i'}^{(\ell)}\boldsymbol{\beta})\right\} \\
&\quad \times \exp\left\{\alpha \boldsymbol{1_p'}\alpha^{-1/2}\frac{1}{\sigma_\beta}\boldsymbol{I_p}\boldsymbol{\beta} - \alpha\boldsymbol{1_p'}\exp\left(\alpha^{-1/2}\frac{1}{\sigma_\beta}\boldsymbol{I_p}\boldsymbol{\beta}\right)\right\} \\
&= \exp\left\{\boldsymbol{\alpha_\beta'}\boldsymbol{H_\beta}\boldsymbol{\beta} - \boldsymbol{\kappa_\beta'}\exp(\boldsymbol{H_\beta}\boldsymbol{\beta})\right\} \\
\boldsymbol{H_\beta} &= \begin{bmatrix} \boldsymbol{X} \\ \alpha^{-1/2}\frac{1}{\sigma_\beta}\boldsymbol{I_p} \end{bmatrix}, \quad \boldsymbol{\alpha_\beta} = (\boldsymbol{\tilde{w}'} \odot \boldsymbol{Z'}, \alpha\boldsymbol{1_p'})', \quad \boldsymbol{\kappa_\beta} = (\boldsymbol{\tilde{w}'} \odot \exp(\boldsymbol{\Psi\eta} + \boldsymbol{\xi})', \alpha\boldsymbol{1_p'})' \\
\boldsymbol{\beta}|\cdot &\sim \text{cMLG}(\boldsymbol{H_\beta}, \boldsymbol{\alpha_\beta}, \boldsymbol{\kappa_\beta})
\end{aligned}
$$

A.1.3 Variance Parameters

$$\begin{aligned}
\frac{1}{\sigma_k}|\cdot &\propto \exp\left\{\alpha\boldsymbol{1_r'}\alpha^{-1/2}\frac{1}{\sigma_k}\boldsymbol{I_r\eta} - \alpha\boldsymbol{1_r'}\text{exp}\left(\alpha^{-1/2}\frac{1}{\sigma_k}\boldsymbol{I_r\eta}\right)\right\} \\
&\times \exp\left\{\omega\frac{1}{\sigma_k} - \rho\exp\left(\frac{1}{\sigma_k}\right)\right\} \times I(\sigma_k > 0) \\
&= \exp\left\{\boldsymbol{\omega_k'H_k}\frac{1}{\sigma_k} - \boldsymbol{\rho_k'}\text{exp}\left(\boldsymbol{H_k}\frac{1}{\sigma_k}\right)\right\} \times I(\sigma_k > 0) \\
&\boldsymbol{H_k} = (\alpha^{-1/2}\boldsymbol{\eta'}, 1)' \quad \boldsymbol{\omega_k} = (\alpha\boldsymbol{1_r'}, \omega)' \quad \boldsymbol{\rho_k} = (\alpha\boldsymbol{1_r'}, \rho)' \\
\frac{1}{\sigma_k}|\cdot &\sim \text{cMLG}(\boldsymbol{H_k, \omega_k, \rho_k}) \times I(\sigma_k > 0)
\end{aligned}$$

$$\begin{aligned}
\frac{1}{\sigma_\xi}|\cdot &\propto \exp\left\{\alpha\boldsymbol{1_n'}\alpha^{-1/2}\frac{1}{\sigma_\xi}\boldsymbol{I_n\xi} - \alpha\boldsymbol{1_n'}\text{exp}\left(\alpha^{-1/2}\frac{1}{\sigma_\xi}\boldsymbol{I_n\xi}\right)\right\} \\
&\times \exp\left\{\omega\frac{1}{\sigma_\xi} - \rho\exp\left(\frac{1}{\sigma_\xi}\right)\right\} \times I(\sigma_\xi > 0) \\
&= \exp\left\{\boldsymbol{\omega_\xi'H_\xi}\frac{1}{\sigma_\xi} - \boldsymbol{\rho_\xi'}\text{exp}\left(\boldsymbol{H_\xi}\frac{1}{\sigma_\xi}\right)\right\} \times I(\sigma_\xi > 0) \\
&\boldsymbol{H_\xi} = (\alpha^{-1/2}\boldsymbol{\xi'}, 1)' \quad \boldsymbol{\omega_\xi} = (\alpha\boldsymbol{1_n'}, \omega)' \quad \boldsymbol{\rho_\xi} = (\alpha\boldsymbol{1_n'}, \rho)' \\
\frac{1}{\sigma_\xi}|\cdot &\sim \text{cMLG}(\boldsymbol{H_\xi, \omega_\xi, \rho_\xi}) \times I(\sigma_\xi > 0)
\end{aligned}$$

A.2 Full Conditional Distributions for Chapter 3

Let $\boldsymbol{\Omega} = \text{diag}(\omega_1, \ldots, \omega_n)$, and $\boldsymbol{\kappa} = (\tilde{w}_1 * (y_1 - n_1/2), \ldots, \tilde{w}_n * (y_n - n_n/2))'$. Note that $\boldsymbol{\kappa}/\boldsymbol{\omega}$ represents element-wise division.

$$
\begin{aligned}
\omega_i|\cdot &\sim \text{PG}(\tilde{w}_i * n_i,\ \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{\psi}_i'\boldsymbol{\eta}),\ i = 1, \ldots, n \\
\boldsymbol{\eta}|\cdot &\propto \prod_{i=1}^{n} \exp\left(\kappa_i \boldsymbol{\phi_i'}\boldsymbol{\eta} - \frac{1}{2}\omega_i(\boldsymbol{\phi_i'}\boldsymbol{\eta})^2 - \omega_i(\boldsymbol{\phi_i'}\boldsymbol{\eta})(\boldsymbol{x_i'}\boldsymbol{\beta})\right) \\
&\quad \times \exp\left(-\frac{1}{2\sigma_\eta^2}\boldsymbol{\eta'}\boldsymbol{\eta}\right) \\
&\propto \exp\left(-\frac{1}{2}(\boldsymbol{\kappa}/\boldsymbol{\omega} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\Phi}\boldsymbol{\eta})'\boldsymbol{\Omega}(\boldsymbol{\kappa}/\boldsymbol{\omega} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\Phi}\boldsymbol{\eta}) - \frac{1}{\sigma_\eta^2}\boldsymbol{\eta'}\boldsymbol{\eta}\right) \\
\boldsymbol{\eta}|\cdot &\sim \text{N}_r\left(\boldsymbol{\mu} = (\boldsymbol{\Phi'}\boldsymbol{\Omega}\boldsymbol{\Phi} + \frac{1}{\sigma_\eta^2}\boldsymbol{I}_r)^{-1}\boldsymbol{\Phi'}\boldsymbol{\Omega}(\boldsymbol{\kappa}/\boldsymbol{\omega} - \boldsymbol{X}\boldsymbol{\beta}),\ \boldsymbol{\Sigma} = (\boldsymbol{\Phi'}\boldsymbol{\Omega}\boldsymbol{\Phi} + \frac{1}{\sigma_\eta^2}\boldsymbol{I}_r)^{-1}\right)
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{\beta}|\cdot &\propto \prod_{i=1}^{n} \exp\left(\kappa_i \boldsymbol{x_i'\beta} - \frac{1}{2}\omega_i(\boldsymbol{x_i'\beta})^2 - \omega_i(\boldsymbol{x_i'\beta})(\boldsymbol{\phi_i'\eta})\right) \\
&\times \exp\left(-\frac{1}{2\sigma_\beta^2}\boldsymbol{\beta'\beta}\right) \\
&\propto \exp\left(-\frac{1}{2}(\boldsymbol{\kappa}/\boldsymbol{\omega} - \boldsymbol{\Phi\eta} - \boldsymbol{X\beta})'\boldsymbol{\Omega}(\boldsymbol{\kappa}/\boldsymbol{\omega} - \boldsymbol{\Phi\eta} - \boldsymbol{X\beta}) - \frac{1}{\sigma_\beta^2}\boldsymbol{\beta'\beta}\right) \\
\boldsymbol{\beta}|\cdot &\sim \mathrm{N}_p\left(\boldsymbol{\mu} = (\boldsymbol{X'\Omega X} + \frac{1}{\sigma_\beta^2}\boldsymbol{I}_p)^{-1}\boldsymbol{X'\Omega}(\boldsymbol{\kappa}/\boldsymbol{\omega} - \boldsymbol{\Phi\eta}),\ \boldsymbol{\Sigma} = (\boldsymbol{X'\Omega X} + \frac{1}{\sigma_\beta^2}\boldsymbol{I}_p)^{-1}\right) \\
\sigma_\eta^2|\cdot &\propto \left(\sigma_\eta^2\right)^{-\frac{r}{2}} \exp\left(-\frac{1}{2\sigma_\eta^2}\boldsymbol{\eta'\eta}\right) \\
&\times \left(\sigma_\eta^2\right)^{-a-1} \exp\left(-\frac{1}{\sigma_\eta^2}b\right) \\
&\propto \left(\sigma_\eta^2\right)^{-(a+\frac{r}{2})-1} \exp\left(-\frac{1}{\sigma_\eta^2}(b + \frac{\boldsymbol{\eta'\eta}}{2})\right) \\
\sigma_\eta^2|\cdot &\sim \mathrm{IG}\left(a + \frac{r}{2},\ b + \frac{\boldsymbol{\eta'\eta}}{2}\right)
\end{aligned}
$$

A.3 Full Conditional Distributions for Chapter 5

Let $\boldsymbol{\Omega} = \mathrm{diag}(\omega_1, \ldots, \omega_n)$, $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1^2, \ldots, \lambda_K^2)$, and $\boldsymbol{\kappa} = (\tilde{w}_1 * (y_1 - n_1/2), \ldots, \tilde{w}_n * (y_n - n_n/2))'$. Note that $\boldsymbol{\kappa}/\boldsymbol{\omega}$ represents element-wise division.

$$\begin{aligned}
\omega_i|\cdot &\sim \mathrm{PG}(\tilde{w}_i * n_i,\ \boldsymbol{x}_i'\boldsymbol{\beta} + \sum_{k=1}^{K} b(k)\xi_i(k)),\ i = 1, \ldots, n \\
\boldsymbol{b}|\cdot &\propto \prod_{i=1}^{n} \exp\left(\kappa_i \boldsymbol{\xi}_i'\boldsymbol{b} - \frac{1}{2}\omega_i(\boldsymbol{\xi}_i'\boldsymbol{b})^2 - \omega_i(\boldsymbol{\xi}_i'\boldsymbol{b})(\boldsymbol{x}_i'\boldsymbol{\beta})\right) \\
&\quad \times \exp\left(-\frac{1}{2\tau^2}\boldsymbol{b}'\boldsymbol{\Lambda}^{-1}\boldsymbol{b}\right) \\
&\quad \propto \exp\left(-\frac{1}{2}(\boldsymbol{\kappa}/\boldsymbol{\omega} - \boldsymbol{X\beta} - \Xi\boldsymbol{b})'\boldsymbol{\Omega}(\boldsymbol{\kappa}/\boldsymbol{\omega} - \boldsymbol{X\beta} - \Xi\boldsymbol{b}) - \frac{1}{2\tau^2}\boldsymbol{b}'\boldsymbol{\Lambda}^{-1}\boldsymbol{b}\right) \\
\boldsymbol{b}|\cdot &\sim \mathrm{N}_K\left(\boldsymbol{\mu} = (\Xi'\boldsymbol{\Omega}\Xi + \frac{1}{\tau^2}\boldsymbol{\Lambda}^{-1})^{-1}\Xi'\boldsymbol{\Omega}(\boldsymbol{\kappa}/\boldsymbol{\omega} - \boldsymbol{X\beta}),\ \boldsymbol{\Sigma} = (\Xi'\boldsymbol{\Omega}\Xi + \frac{1}{\tau^2}\boldsymbol{\Lambda}^{-1})^{-1}\right)
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{\beta}|\cdot &\propto \prod_{i=1}^{n} \exp\left(\kappa_i \boldsymbol{x}_i'\boldsymbol{\beta} - \frac{1}{2}\omega_i(\boldsymbol{x}_i'\boldsymbol{\beta})^2 - \omega_i(\boldsymbol{x}_i'\boldsymbol{\beta})(\boldsymbol{\xi}_i'\boldsymbol{b})\right) \\
&\quad \times \exp\left(-\frac{1}{2\sigma_\beta^2}\boldsymbol{\beta}'\boldsymbol{\beta}\right) \\
&\quad \propto \exp\left(-\frac{1}{2}(\boldsymbol{\kappa}/\boldsymbol{\omega} - \Xi\boldsymbol{b} - \boldsymbol{X\beta})'\boldsymbol{\Omega}(\boldsymbol{\kappa}/\boldsymbol{\omega} - \Xi\boldsymbol{b} - \boldsymbol{X\beta}) - \frac{1}{2\sigma_\beta^2}\boldsymbol{\beta}'\boldsymbol{\beta}\right) \\
\boldsymbol{\beta}|\cdot &\sim \mathrm{N}_q\left(\boldsymbol{\mu} = (\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X} + \frac{1}{\sigma_\beta^2}\boldsymbol{I}_q)^{-1}\boldsymbol{X}'\boldsymbol{\Omega}(\boldsymbol{\kappa}/\boldsymbol{\omega} - \Xi\boldsymbol{b}),\ \boldsymbol{\Sigma} = (\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X} + \frac{1}{\sigma_\beta^2}\boldsymbol{I}_q)^{-1}\right)
\end{aligned}$$

$$
\begin{aligned}
\lambda_k^2|\cdot &\propto (\lambda_k^2)^{-1/2}\text{exp}\left(-\frac{b(k)^2}{2\tau^2\lambda_k^2}\right)\\
&\quad\times(\lambda_k^2)^{-3/2}\text{exp}\left(-\frac{1}{\nu_k\lambda_k^2}\right)\\
&\quad\propto(\lambda_k^2)^{-2}\text{exp}\left\{-\frac{1}{\lambda_k^2}\left(\frac{1}{\nu_k}+\frac{b(k)^2}{2\tau^2}\right)\right\}\\
\lambda_k^2|\cdot &\sim \text{IG}\left(1,\frac{1}{\nu_k}+\frac{b(k)^2}{2\tau^2}\right)
\end{aligned}
$$

$$
\begin{aligned}
\tau^2|\cdot &\propto (\tau^2)^{-K/2}\text{exp}\left(-\frac{1}{\tau^2}\sum_{k=1}^{K}\frac{b(k)^2}{2\lambda_k^2}\right)\\
&\quad\times(\tau^2)^{-3/2}\text{exp}\left(-\frac{1}{\nu_\tau\tau^2}\right)\\
&\quad\propto(\tau^2)^{-\frac{K+1}{2}-1}\text{exp}\left\{-\frac{1}{\tau^2}\left(\frac{1}{\nu_\tau}+\sum_{k=1}^{K}\frac{b(k)^2}{2\lambda_k^2}\right)\right\}\\
\tau^2|\cdot &\sim \text{IG}\left(\frac{K+1}{2},\frac{1}{\nu_\tau}+\sum_{k=1}^{K}\frac{b(k)^2}{2\lambda_k^2}\right)
\end{aligned}
$$

$$
\begin{aligned}
\nu_k|\cdot &\propto \nu_k^{-3/2}\text{exp}\left(-\frac{1}{\nu_k}\right)\\
&\quad\times\nu_k^{-1/2}\text{exp}\left(-\frac{1}{\nu_k\lambda_k^2}\right)\\
&\quad\propto\nu_k^{-2}\text{exp}\left\{-\frac{1}{\nu_k}\left(1+\frac{1}{\lambda_k^2}\right)\right\}\\
\nu_k|\cdot &\sim \text{IG}\left(1,1+\frac{1}{\lambda_k^2}\right)
\end{aligned}
$$

$$
\begin{aligned}
\nu_\tau|\cdot &\propto \nu_\tau^{-3/2}\text{exp}\left(-\frac{1}{\nu_\tau}\right)\\
&\quad\times\nu_\tau^{-1/2}\text{exp}\left(-\frac{1}{\nu_\tau\tau^2}\right)\\
&\quad\propto\nu_\tau^{-2}\text{exp}\left\{-\frac{1}{\nu_\tau}\left(1+\frac{1}{\tau^2}\right)\right\}\\
\nu_\tau|\cdot &\sim \text{IG}\left(1,1+\frac{1}{\tau^2}\right)
\end{aligned}
$$

Appendix B

Poststratification

B.1 Poststratification with a Variational Distribution for Chapter 3

To construct our estimates, we require response predictions for every unit in the population. We will assume for the sake of exposition that the responses are binary, but these same techniques can be applied to categorical responses as well.

Because our model utilizes only categorical rather than continuous covariates, computation can be simplified through the use of poststratification cells. Specifically, let $j = 1, \dots, J$ index the J unique poststratification cells (e.g. the unique combinations of categorical covariates and county indicators). Each cell is also associated with a population size, N_j. Within each cell, population units are exchangeable, and predicted responses can be generated from the same distribution. To estimate p_j, the probability of a successful outcome in cell j, we require estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ as well as the vector of cell covariates, $\boldsymbol{x}_j$ and the vector of spatial basis functions for the cell, $\boldsymbol{\phi}_j$.

To begin, we work with our variational distribution for $\boldsymbol{\zeta} = (\boldsymbol{\beta'}, \boldsymbol{\eta'})$. We can sample from this distribution by sampling from a $\mathrm{N}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ distribution, where $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$ are estimated from the variational Bayes procedure outlined in Algorithm 1. We take R total posterior samples, yielding $\boldsymbol{\beta}^{(r)}$ and $\boldsymbol{\eta}^{(r)}$ for $r = 1, \ldots, R$. Then, for each posterior sample r and each cell j, we generate the population (i.e. the number of positive responses) within the cell,$s_j^{(r)}$, by sampling from $\mathrm{Bin}(N_j, p_j^{(r)})$, where $p_j^{(r)} = \mathrm{logit}^{-1}(\boldsymbol{x}_j'\boldsymbol{\beta}^{(r)} + \boldsymbol{\phi}_j'\boldsymbol{\eta}^{(r)})$. Having effectively generated a synthetic population, we can aggregate the units within a given domain to generate a population estimate. For example, for a given iteration r, we can create an estimate of the population proportion in county c as

$$p_c^{(r)} = \frac{\sum_{j \in c} s_j^{(r)}}{\sum_{j \in c} N_j}.$$

Then, for our point estimate of the population proportion in county c, we use the posterior mean,

$$\hat{p}_c = \frac{1}{R} \sum_{r=1}^{R} p_c^{(r)}.$$

References

Albert, J. H. and Chib, S. (1993). “Bayesian analysis of binary and polychotomous response data.” *Journal of the American Statistical Association*, 88, 422, 669–679.

Asparouhov, T. (2006). “General multi-level modeling with sampling weights.” *Communications in Statistics—Theory and Methods*, 35, 3, 439–460.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). “Gaussian predictive process models for large spatial data sets.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 4, 825–848.

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). “An error-components model for prediction of county crop areas using survey and satellite data.” *Journal of the American Statistical Association*, 83, 401, 28–36.

Bauder, M., Luery, D., and Szelepka, S. (2018). “Small area estimation of health insurance coverage in 2010 – 2016.” Tech. rep., Small Area Methods Branch, Social, Economic, and Housing Statistics Division, U. S. Census Bureau.

Beal, M. J. and Ghahramani, Z. (2003). “The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures.” In

Bayesian Statistics, eds. J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, vol. 7, 453–464. Oxford, UK: Oxford University Press.

Beaumont, J.-F. (2008). “A new approach to weighting and inference in sample surveys.” *Biometrika*, 95, 3, 539–553.

Besag, J. (1974). “Spatial interaction and the statistical analysis of lattice systems (with discussion).” *Journal of the Royal Statistical Society. Series B*, 36, 192 – 236.

Besag, J., York, J., and Mollié, A. (1991). “Bayesian image restoration, with two applications in spatial statistics.” *Annals of the Institute of Statistical Mathematics*, 43, 1, 1–20.

Binder, D. A. (1983). “On the variances of asymptotically normal estimators from complex surveys.” *International Statistical Review*, 51, 3, 279–292.

Binder, D. A. and Patak, Z. (1994). “Use of estimating functions for estimation from complex surveys.” *Journal of the American Statistical Association*, 89, 427, 1035–1043.

Bingham, E. and Mannila, H. (2001). “Random projection in dimensionality reduction: applications to image and text data.” In *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 245–250. ACM.

Bradley, J. R., Cressie, N., and Shi, T. (2016). “A comparison of spatial predictors when datasets could be very large.” *Statistics Surveys*, 10, 100–131.

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2015). "Multivariate spatio-temporal models for high-dimensional areal data with application to Longitudinal Employer-Household Dynamics." *The Annals of Applied Statistics*, 9, 4, 1761–1791.

— (2018). "Computationally Efficient Multivariate Spatio-Temporal Models for High-Dimensional Count-Valued Data (with Discussion)." *Bayesian Analysis*, 13, 1, 253–310.

— (2020). "Bayesian hierarchical models with conjugate full-conditional distributions for dependent data from the natural exponential family." *Journal of the American Statistical Association*, 115, 532, 2037–2052.

Brewer, K., Early, L., and Hanif, M. (1984). "Poisson, modified Poisson and collocated sampling." *Journal of Statistical Planning and Inference*, 10, 1, 15–30.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). "Stan: A probabilistic programming language." *Journal of Statistical Software*, 76, 1.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97, 2, 465–480.

Chamara, L., Zhou, H., Huang, G., and Vong, C. (2013). "Representational learning with extreme learning machine for big data." *IEEE Intelligent Systems*, 28, 6, 31–34.

Chen, C., Wakefield, J., and Lumely, T. (2014). "The use of sampling weights in Bayesian hierarchical models for small area estimation." *Spatial and Spatio-temporal Epidemiology*, 11, 33–43.

Chen, Y., Yang, J., Wang, C., and Park, D. (2016). "Variational Bayesian extreme learning machine." *Neural Computing and Applications*, 27, 1, 185–196.

Congdon, P. and Lloyd, P. (2010). "Estimating small area diabetes prevalence in the US using the behavioral risk factor surveillance system." *Journal of Data Science*, 8, 2, 235–252.

Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. John Wiley & Sons.

DeBell, M. (2013). "Harder than it looks: Coding political knowledge on the ANES." *Political Analysis*, 393–406.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). "Model-based geostatistics." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47, 3, 299–350.

Dong, Q., Elliott, M. R., and Raghunathan, T. E. (2014). "A nonparametric method to generate synthetic populations to adjust for complex sampling design features." *Survey Methodology*, 40, 1, 29–46.

Durante, D., Rigon, T., et al. (2019). "Conditionally conjugate mean-field variational Bayes for logistic models." *Statistical Science*, 34, 3, 472–485.

Fay, R. E. and Herriot, R. A. (1979). "Estimates of income for small places: an application of James-Stein procedures to census data." *Journal of the American Statistical Association*, 74, 366a, 269–277.

Franco, C. and Bell, W. R. (2014). "Borrowing information overtime in binomial/logit normal models for small area estimation." *Statistics in Transition*, 16, 563 – 584.

Gelfand, A. E. and Schliep, E. M. (2016). “Spatial statistics and Gaussian processes: A beautiful marriage.” *Spatial Statistics*, 18, 86–104.

Gelman, A. (2007). “Struggles with survey weighting and regression modeling.” *Statistical Science*, 22, 2, 153–164.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. London: Chapman and Hall.

Gelman, A. and Little, T. C. (1997). “Poststratification into many categories using hierarchical logistic regression.” *Survey Methodology*, 23, 127–135.

Geweke, J. F. (1991). “Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments.” Staff Report 148, Federal Reserve Bank of Minneapolis.

Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Di, C., Gellar, J., Harezlak, J., McLean, M. W., Swihart, B., Xiao, L., Crainiceanu, C., and Reiss, P. T. (2019). *refund: Regression with Functional Data*. R package version 0.1-21.

Grilli, L. and Pratesi, M. (2004). “Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs.” *Survey Methodology*, 30, 1, 93–103.

Guadarrama, M., Molina, I., and Rao, J. N. K. (2018). “Small area estimation of general parameters under complex sampling designs.” *Computational Statistics and Data Analysis*, 121, 20 – 40.

Hidiroglou, M. A. and You, Y. (2016). “Comparison of unit level and area level small area estimators.” *Survey Methodology*, 42, 41–61.

Holan, S. H., Wikle, C. K., Sullivan-Beckers, L. E., and Cocroft, R. B. (2010). "Modeling complex phenotypes: generalized linear models using spectrogram predictors of animal communication signals." *Biometrics*, 66, 3, 914–924.

Horvitz, D. G. and Thompson, D. J. (1952). "A generalization of sampling without replacement from a finite universe." *Journal of the American Statistical Association*, 47, 663 – 685.

Horwitz, R., Brockhaus, S., Henninger, F., Kieslich, P. J., Schierholz, M., Keusch, F., and Kreuter, F. (2020). "Learning from mouse movements: Improving questionnaires and respondents' user experience through passive data collection." *Advances in Questionnaire Design, Development, Evaluation and Testing*, 403–425.

Hsing, T. and Eubank, R. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*, vol. 997. John Wiley & Sons.

Huang, G., Huang, G.-B., Song, S., and You, K. (2015). "Trends in extreme learning machines: A review." *Neural Networks*, 61, 32–48.

Huang, G.-B., Zhu, Q.-Y., and Siew, C.-K. (2006). "Extreme learning machine: theory and applications." *Neurocomputing*, 70, 1-3, 489–501.

Hughes, J. and Haran, M. (2013). "Dimension reduction and alleviation of confounding for spatial generalized linear mixed models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75, 1, 139–159.

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

Jin, X., Carlin, B. P., and Banerjee, S. (2005). "Generalized hierarchical multivariate CAR models for areal data." *Biometrics*, 61, 4, 950–961.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). "An introduction to variational methods for graphical models." *Machine Learning*, 37, 2, 183–233.

Kim, D. H. (2002). "Bayesian and empirical Bayesian analysis under informative sampling." *Sankhyā: The Indian Journal of Statistics, Series B*, 64, 3, 267–288.

Kim, J. K., Park, S., and Lee, Y. (2017). "Statistical inference using generalized linear mixed models under informative cluster sampling." *Canadian Journal of Statistics*, 45, 4, 479–497.

Kim, S., Shephard, N., and Chib, S. (1998). "Stochastic volatility: likelihood inference and comparison with ARCH models." *The Review of Economic Studies*, 65, 3, 361–393.

Kish, L. (1965). *Survey Sampling*. New York: Wiley.

— (1992). "Weighting for unequal P$_i$." *Journal of Official Statistics*, 8, 2, 183.

Kokoszka, P. and Reimherr, M. (2017). *Introduction to functional data analysis*. Chapman and Hall/CRC.

Korn, E. L. and Graubard, B. I. (2003). "Estimating variance components by using survey data." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65, 1, 175–190.

León-Novelo, L. G., Savitsky, T. D., et al. (2019). "Fully Bayesian estimation under informative sampling." *Electronic Journal of Statistics*, 13, 1, 1608–1645.

Leroux, A., Di, J., Smirnova, E., Mcguffey, E. J., Cao, Q., Bayatmokhtari, E., Tabacu, L., Zipunnikov, V., Urbanek, J. K., and Crainiceanu, C. (2019). "Organizing and analyzing the activity data in nhanes." *Statistics in Biosciences*, 11, 2, 262–287.

Linderman, S., Johnson, M. J., and Adams, R. P. (2015). "Dependent multinomial models made easy: Stick-breaking with the Pólya-Gamma augmentation." In *Advances in Neural Information Processing Systems*, 3456–3464.

Little, R. J. (1993). "Post-stratification: a modeler's perspective." *Journal of the American Statistical Association*, 88, 423, 1001–1012.

— (2012). "Calibrated Bayes, an alternative inferential paradigm for official statistics." *Journal of Official Statistics*, 28, 3, 309.

Liu, J. S. (1994). "The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem." *Journal of the American Statistical Association*, 89, 427, 958–966.

Lohr, S. (2010). *Sampling: Design and Analysis, 2nd ed*. Boston: Brooks/Cole.

Luery, D. M. (2011). "Small area income and poverty estimates program." In *Proceedings of 27th SCORUS Conference*, 93–107.

Lumley, T. and Scott, A. (2017). "Fitting regression models to survey data." *Statistical Science*, 32, 2, 265 – 278.

Makalic, E. and Schmidt, D. F. (2015). “A simple sampler for the horseshoe estimator.” *IEEE Signal Processing Letters*, 23, 1, 179–182.

Malec, D., Davis, W. W., and Cao, X. (1999). “Model-based small area estimates of overweight prevalence using sample selection adjustment.” *Statistics in Medicine*, 18, 23, 3189–3200.

Malec, D., Sedransk, J., Moriarity, C. L., and LeClere, F. B. (1997). “Small area inference for binary variables in the National Health Interview Survey.” *Journal of the American Statistical Association*, 92, 439, 815–826.

McConville, K., Tang, B., Zhu, G., Cheung, S., and Li, S. (2018). *mase: Model-Assisted Survey Estimation*.

McDermott, P. L. and Wikle, C. K. (2017). “An ensemble quadratic echo state network for non-linear spatio-temporal forecasting.” *Stat*, 6, 1, 315–330.

— (2019). “Bayesian recurrent neural network models for forecasting and quantifying uncertainty in spatial-temporal data.” *Entropy*, 21, 2, 184.

Midzuno, H. (1951). “On the sampling system with probability proportionate to sum of sizes.” *Annals of the Institute of Statistical Mathematics*, 3, 1, 99–107.

Molina, I., Nandram, B., and Rao, J. (2014). “Small area estimation of general parameters with application to poverty indicators: a hierarchical Bayes approach.” *The Annals of Applied Statistics*, 8, 2, 852–885.

Morris, J. S. (2015). “Functional regression.” *Annual Review of Statistics and Its Application*, 2, 321–359.

Nathan, G. and Holt, D. (1980). "The effect of survey design on regression analysis." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 42, 3, 377–386.

Park, D. K., Gelman, A., and Bafumi, J. (2004). "Bayesian multilevel estimation with poststratification: State-level estimates from national polls." *Political Analysis*, 375–385.

— (2006). "State-level opinions from national surveys: Poststratification using multilevel logistic regression." *Public Opinion in State Politics*.

Pfeffermann, D., Krieger, A. M., and Rinott, Y. (1998a). "Parametric distributions of complex survey data under informative probability sampling." *Statistica Sinica*, 8, 1087–1114.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., and Rasbash, J. (1998b). "Weighting for unequal selection probabilities in multilevel models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60, 1, 23–40.

Pfeffermann, D. and Sverchkov, M. (1999). "Parametric and semi-parametric estimation of regression models fitted to survey data." *Sankhyā: The Indian Journal of Statistics, Series B*, 61, 166–186.

— (2007). "Small-area estimation under informative probability sampling of areas and within the selected areas." *Journal of the American Statistical Association*, 102, 480, 1427–1439.

Polson, N. G., Scott, J. G., and Windle, J. (2013). "Bayesian inference for logistic models using Pólya–Gamma latent variables." *Journal of the American Statistical Association*, 108, 504, 1339–1349.

Porter, A. T., Holan, S. H., Wikle, C. K., and Cressie, N. (2014). "Spatial Fay–Herriot models for small area estimation with functional covariates." *Spatial Statistics*, 10, 27–42.

Porter, A. T., Wikle, C. K., and Holan, S. H. (2015). "Small area estimation via multivariate Fay–Herriot models with latent spatial dependence." *Australian & New Zealand Journal of Statistics*, 57, 1, 15–29.

Potthoff, R. F., Woodbury, M. A., and Manton, K. G. (1992). ""Equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey weights under superpopulation models." *Journal of the American Statistical Association*, 87, 418, 383–396.

Prasad, N. and Rao, J. (1990). "The estimation of mean squared error of small-area estimators." *Journal of the American Statistical Association*, 85, 163 – 171.

Prokhorov, D. (2005). "Echo state networks: appeal and challenges." In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, vol. 3, 1463–1466. IEEE.

Rabe-Hesketh, S. and Skrondal, A. (2006). "Multilevel modelling of complex survey data." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169, 4, 805–827.

Ramsay, J. and Silverman, B. W. (2005). *Functional data analysis*. Springer.

Rao, J., Verret, F., and Hidiroglou, M. A. (2013). "A weighted composite likelihood approach to inference for two-level models from survey data." *Survey Methodology*, 39, 2, 263–282.

Rao, J. and Wu, C. (2010). "Bayesian pseudo-empirical-likelihood intervals for complex surveys." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72, 4, 533–544.

Rao, J. N. K. and Molina, I. (2015). *Small Area Estimation*. Hoboken, New Jersey: Wiley.

Ribatet, M., Cooley, D., and Davison, A. C. (2012). "Bayesian inference from composite likelihoods, with an application to spatial extremes." *Statistica Sinica*, 22, 813–845.

Savitsky, T. D. and Toth, D. (2016). "Bayesian estimation under informative sampling." *Electronic Journal of Statistics*, 10, 1, 1677–1708.

Schuna, J. M., Johnson, W. D., and Tudor-Locke, C. (2013). "Adult self-reported and objectively monitored physical activity and sedentary behavior: NHANES 2005–2006." *International Journal of Behavioral Nutrition and Physical Activity*, 10, 1, 126.

Si, Y., Pillai, N. S., Gelman, A., et al. (2015). "Bayesian nonparametric weighted sampling inference." *Bayesian Analysis*, 10, 3, 605–625.

Skinner, C. J. (1989). "Domain means, regression and multivariate analysis." In *Analysis of Complex Surveys*, eds. C. J. Skinner, D. Holt, and T. M. F. Smith, chap. 2, 80 – 84. Chichester: Wiley.

Soria-Olivas, E., Gomez-Sanchis, J., Martin, J. D., Vila-Frances, J., Martinez, M., Magdalena, J. R., and Serrano, A. J. (2011). “BELM: Bayesian extreme learning machine.” *IEEE Transactions on Neural Networks*, 22, 3, 505–509.

Stan Development Team (2021). “Stan modeling language users guide and reference manual, version 2.26.” https://mc-stan.org.

Tillé, Y. and Matei, A. (2016). *sampling: Survey Sampling*. R package version 2.8.

Vandendijck, Y., Faes, C., Kirby, R. S., Lawson, A., and Hens, N. (2016). “Model-based inference for small area estimation with sampling weights.” *Spatial Statistics*, 18, 455–473.

Varin, C., Reid, N., and Firth, D. (2011). “An overview of composite likelihood methods.” *Statistica Sinica*, 21, 5–42.

Verret, F., Rao, J. N. K., and Hidiroglou, M. A. (2015). “Model-based small area estimation under informative sampling.” *Survey Methodology*, 41, 333 – 347.

Wainwright, M. J., Jordan, M. I., et al. (2008). “Graphical models, exponential families, and variational inference.” *Foundations and Trends® in Machine Learning*, 1, 1–2, 1–305.

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016). “Functional data analysis.” *Annual Review of Statistics and Its Application*, 3, 257–295.

Windle, J., Polson, N., and Scott, J. (2013). “BayesLogit: Bayesian logistic regression.” http://cran.r-project.org/web/packages/BayesLogit/index.html. R package version 0.2-4.

Xiao, L., Zipunnikov, V., Ruppert, D., and Crainiceanu, C. (2016). “Fast covariance estimation for high-dimensional functional data.” *Statistics and Computing*, 26, 1-2, 409–421.

Yang, W.-H., Wikle, C. K., Holan, S. H., and Wildhaber, M. L. (2013). “Ecological prediction with nonlinear multivariate time-frequency functional data models.” *Journal of Agricultural, Biological, and Environmental Statistics*, 18, 3, 450–474.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005). “Functional data analysis for sparse longitudinal data.” *Journal of the American Statistical Association*, 100, 470, 577–590.

Yi, G., Rao, J. N. K., and Li, H. (2016). “A weighted composite likelihood approach for analysis of survey data under two-level models.” *Statistica Sinica*, 26, 569 – 587.

You, Y. and Rao, J. (2002). “A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights.” *Canadian Journal of Statistics*, 30, 3, 431–439.

Zhang, X., Holt, J. B., Lu, H., Wheaton, A. G., Ford, E. S., Greenlund, K. J., and Croft, J. B. (2014). “Multilevel regression and poststratification for small-area estimation of population health outcomes: a case study of chronic obstructive pulmonary disease prevalence using the behavioral risk factor surveillance system.” *American Journal of Epidemiology*, 179, 8, 1025–1033.

Zheng, H. and Little, R. J. (2003). “Penalized spline model-based estimation of the finite populations total from probability-proportional-to-size samples." *Journal of Official Statistics*, 19, 2, 99.

VITA

Paul A. Parker was born in Boise, Idaho, on March 4, 1992. After graduating from Borah High School, he moved to Moscow, Idaho to attend the University of Idaho. After four years, he graduated in 2014 with a Bachelor of Science in Mathematics. After working as a data analyst in Seattle, Washington for two years, he moved to Columbia, Missouri to begin graduate studies in August 2016. In January 2017 he began working with Scott H. Holan as his dissertation advisor and received his Doctor of Philosophy in Statistics in August 2021.

Paul has been happily married to Mackenzie M. Parker since August 2017. Paul will begin his academic career in the fall of 2021 as an assistant professor in the Department of Statistics at the University of California Santa Cruz.