Bayesian hierarchical weighting adjustment and survey inference

Yajuan Si[*], Rob Trangucci, Jonah Sol Gabry and Andrew Gelman[†]

19 July 2017

Abstract

We combine Bayesian prediction and weighted inference as a unified approach to survey inference. The general principles of Bayesian analysis imply that models for survey outcomes should be conditional on all variables that affect the probability of inclusion. We incorporate the weighting variables under the framework of multilevel regression and poststratification, and as a byproduct generating model-based weights after smoothing. We investigate deep interactions and introduce structured prior distributions for smoothing and stability of estimates. The computation is done via Stan and implemented in an open source R package *Rstanarm* ready for public use. Simulation studies illustrate that model-based prediction and weighting inference outperform classical weighting. We apply the proposal to the New York Longitudinal Study of Wellbeing. The new approach generates robust weights and increases efficiency for finite population inference, especially for subsets of the population.

1. Introduction

Survey data can be unrepresentative of the underlying population due to planned design features and unplanned nonresponse or undercoverage issues. Arguments between design-based and model-based inference have a long history in survey research (Little, 2004). The former automatically takes into account of survey design, while the latter can yield robust inference for small sample estimation. Design-based approaches use weights to balance the sample and the population; see Chen et al. (2017) for a review of various weighted estimators for a survey population mean. However, classical survey weighting usually relies on many user-defined choices such that the process of weighting is difficult to codify (Gelman, 2007). The Bayesian approach for finite population inference (Ghosh and Meeden, 1997) allows prior information to be incorporated, when appropriate, but is subject to model misspecification.

In the present paper we combine Bayesian prediction and weighted inference as a unified approach to survey inference, applying scalable and robust Bayesian regression models to account for complex design features under the framework of multilevel regression and poststratification (MRP, Gelman and Little (1997); Park et al. (2005); Ghitza and Gelman (2013)). The proposal yields design-consistent and efficient finite population inference, especially for subgroups, and constructs model-based weights after smoothing.

For a finite population of N units, we denote the variable of interest as $y = (y_1, \ldots, y_N)$ and the inclusion indicator variable as $I = (I_1, \ldots, I_N)$, where $I_i = 1$ if unit i is included in the sample and $I_i = 0$ otherwise. Here, inclusion refers to selection and response. Design-based inference considers the distribution of I and treats y as fixed. Model-based inference considers the joint distribution for I and y.

To account for the factors that affect inclusion, design weights adjust for unequal probabilities of sampling, and the subsequent weighting accounts for coverage problems and nonresponse during data collection or data cleaning. Classical weights are thus generated as a product of multiple

[*]University of Wisconsin-Madison, corresponding email: ysi@biostat.wisc.edu
[†]Columbia University

adjustment factors: inverse probability of selection, inverse propensity score of response, and post-stratification (Holt and Smith, 1979) (also called calibration, benchmarking) ratios. Each of these adjustments can be approximate when probability of selection, probability of response, or population totals are estimated from data. Beyond any approximation issues, even if the inclusion model is known exactly, extreme values of weights will cause high variability and then inferential problems, especially when the weights are weakly correlated with the survey outcome variable. When the weighting process involves poststratification or nonresponse adjustment—where the weights themselves are random variables—the variance estimation will be different from the cases only with fixed design weights. It is non-trivial to derive a variance estimator under multi-stage weighting adjustment or complex sampling design. Current variance estimation approaches such as Taylor approximation and resampling methods are in lack of rigorous justification or evaluation.

In practice, the construction of survey weights requires somewhat arbitrary decisions of the selection of variables and interactions, pooling of weighting cells, and weight trimming. It is unclear whether and how to incorporate auxiliary information (Groves and Couper, 1995). Discussion of smoothing and trimming in the survey weighting literature (e.g. Potter, 1988, 1990; Elliott and Little, 2000; Elliott, 2007; Xia and Elliott, 2016) has focused on estimating the finite population total or mean, with less attention to subdomain estimates. Beaumont (2008) proposes to smooth weights by predicting and regressing these on the survey variables, where the direction is inspiring but tangential to the inference objective. Borrowing information on survey outcomes potentially increases efficiency and calls for a general framework.

Under probability sampling, model-based inferences can be based on the distribution of y alone given the weighting variables are included in the model (Rubin, 1976, 1983). The inclusion mechanism is ignorable when the distribution of I given y is independent of the distribution of y conditional on the weighting variables. Gelman (2007) recommends regression models including weighting variables as covariates. Any of these approaches can be sensitive to prior specification for stable estimation; this is the model-based counterpart to the decisions required for smoothing or trimming classical survey weights.

Model-based and model-assisted weighting adjustment methods for finite population total estimation have been compared by Henry and Valliant (2012). The model-based weighting methods in the superpopulation perspective (Valliant et al., 2000) use predictions from regression models to derive case weights, where the predictions are based on hierarchical linear regression models with various bias corrections (Chambers et al., 1993; Firth and Bennett, 1998). The model-assisted methods derive case weights mainly from calibration on benchmark variables (Kott, 2009) via the generalized regression estimator (GREG, Deville and Särndal (1992)) and penalized spline regression estimators (Breidt et al., 2005). However, the case weights derived from regression predictions can be highly variable and even negative, and may damage some domain estimates.

To protect against model misspecification, Little (1983) recommends modeling differences in the distribution of outcomes across classes defined by differential probabilities of inclusion. Si et al. (2015) construct poststratification cells based on the unique values of inclusion probabilities and build hierarchical models to smooth cell estimates as advocated by Little (1991, 1993).

We propose to use Bayesian hierarchical models accounting for survey design to generate weights that can be used in design-based inference. The inference is well calibrated and valid with good frequentist properties (Little, 2011). For large samples, the inference will parallel with design-based inference. For small samples, the hierarchical model smoothing will stabilize domain estimation and generate robust weighting adjustment.

We use the intrinsic weighting variables, assume they are discretized and construct poststratification cells based on the cross-tabulation. Weights are derived through the regressing survey outcome on weighting variables given the poststratification. The inclusion of the outcome variable into weighting and poststratification both avoid model misspecification and potentially increase efficiency (Fuller, 2009). Multilevel model estimates shrink the cell estimates towards the prediction from the regression model. The MRP framework combines multilevel regression and poststratification, accounts for design features in the Bayesian paradigm, and is then well equipped to handle complex design features. Our proposal distinguishes from the model-based weights in the literature by using the poststratification cell structure and improves by smoothing, thus avoiding negative weight values.

Si et al. (2015) incorporate weights into MRP, increasing flexibility and efficiency comparing to the pseudo-likelihood approach (Pfeffermann, 1993). In the present paper we go further, starting from the variables that are used for weighting and constructing model-based weights as byproducts under MRP. We develop a novel prior specification for the regularization to handle potentially large numbers of poststratification cells. The prior setting allows for variable selection and keeps the hierarchical structure among main effects and high-order interaction terms for categorical variables. That is, if one variable is not predictive, then the high-order interactions involved with this variable are also likely to be not predictive, to facilitate model interpretation.

We have implemented the computation in a R package *rstanarm* (Goodrich and Gabry, 2017) released on CRAN. The fully Bayesian inference is realized via Stan (Stan Development Team, 2017a,b), which uses Hamiltonian Monte Carlo sampler with adaptive path lengths (Hoffman and Gelman, 2014). Stan promotes robust model-based approaches by reducing the computational burden of building and testing new models. The *rstanarm* package allows for efficient Bayesian hierarchical modeling and weighting inference. The codes are publicly available and reproducible. Our developed computation software provide the accessible platform and has the potential to support the unified framework for survey inference.

Section 2 introduces the motivating problem to construct weights for an ongoing social science survey. We discuss the detail of our proposed method in Section 3. Section 4 describes the statistical evaluation on model-based prediction and weighting inference, and demonstrate the efficiency agains in comparison with classical weighting. We apply the proposal to the real-life survey in Section 5. Section 6 summarizes the improvement and discusses further extension.

2. The motivating application

Our methodological research is motivated by operational weighting practice for ongoing surveys. Our immediate goal is to construct weights for the New York City Longitudinal Study of Wellbeing (LSW) (Si and Gelman, 2014; Wimer et al., 2014), a survey organized by the Columbia University Population Research Center, aiming to provide assessments of income poverty, material hardship, and child and family wellbeing of NYC residents.

We use the LSW as an example to illustrate practical weighting issues and our proposed improvement, with the understanding that similar concerns arise in other surveys. The sample includes a phone sample based on random digit dialing and an in-person respondent driven sample of beneficiaries from Robin Hood philanthropic services and their acquaintances. We focus on the phone survey here as an illustration. The LSW phone survey interviews 2,002 NYC adult residents, including 500 cell phone calls and 1502 landline telephone calls, where half of the landline samples

are from low-income NYC areas defined by zipcode information. The collected baseline samples are followed up every three months. We match the samples to the 2011 American Community Survey (ACS) records for NYC. The discrepancies are mainly caused by the oversampling of the low-income neighborhoods and nonresponse.

The baseline weighting process (Si and Gelman, 2014) adjusts for the unequal probability of selection, coverage bias, and nonresponse. Classical weights are products of estimated inverse probability of inclusion and raking (Deville et al., 1993) ratios. However, practitioners have to make arbitrary or subjective choices on the selection and values of weighting factors. For example, to construct weights for individual adults, we have to weight up respondents from large households, as just one adult per sampled household is included in the sample. Gelman and Little (1998) recommend the square root of the ratio of household sizes to family sizes for this weighting adjustment because using household sizes as weights (for example, ACS Weighting Method, 2014) tend to overcorrect in telephone surveys. The raking operation procedure in practice adjusts for socio-demographic factors without tailoring for particular surveys.

The survey organizers are interested in the aspects of life quality of city residents, such as the percentage of children who live under poverty and material hardship. Thus it is important to get accurate estimates for subpopulations. We would like to develop an objective procedure and let the collected survey data determine the weighting process. The basic principle is to adjust for all variables that could affect the selection and response into weighting. Ideally, we expect that weighting variables should include phone availability (number of landline/cell phones and duration with interrupted service), family structure, household structure, socio-demographics and potentially their high-order interaction terms. However, the ACS records only provide information on family size, age, ethnicity, sex, education and poverty gap (a family poverty measure). Meanwhile, considering the substantive analysis goal, the variables describing the number of elder people in the family, the number of children in the family, and the family size, as well as their interactions with poverty gap are recommended by the survey organizers to be included into the weighting process to balance the distribution discrepancy with the population.

To generate classical weights, we select the raking factors that could affect the selection and response, including sex, age, education, ethnicity, poverty gap, the number of children in the family, the number of elder people in the family, the number of working aged people in the family, the two-way interaction between age and poverty gap, the two-way interaction between the number of persons in the family and poverty gap, the two-way interaction between the number of children in the family and poverty gap, and the two-way interaction between the number of elder people in the family and poverty gap. We collect the marginal distributions from ACS and implement raking adjustment. The generated weights have to be trimmed due to some extreme values.

However, it is possible that the subjective weighting adjustment includes some variables or interactions that are not essentially predictive or does not take account for all the important factors that could be of substantive interest later. The raking adjustment assumes that these factors are independent. This will cause biased domain inference if the correlation structure in the sample is different from that in the population. Ideally, we should match based on the joint distribution of these weighting related variables. However, small cell sizes or empty under the deep interactions will lead to extremely large weights that need cell collapsing.

The problems about classical weighting faced by the LSW baseline survey are general enough that they occur for most operational weighting practice in real-life surveys, which could be complicated with complex design, longitudinal structure or multi-stage response mechanisms. Usually

different weighting practitioners generate different weights and then weighted analyses, causing weighting a mess. It is important to propose a model-based weighting procedure that allows the data to select weighting factors without arbitrary choices and facilitates domain estimates. We would like to incorporate these weighting variables into the model for survey outcomes for efficiency gains, model their high-order interaction terms under regularized prior setting and generate the weights that can be equally treated as classical weights. Large number of weighting variables and deep interactions will cause small weighting cells based on the cross-tabulation. The small weighting cells call for statistical adjustment for smoothness and stability.

MRP have achieved success for domain estimation at much finer levels. Borrowing the strength of hierarchical modeling framework with informative prior distribution, we should be able to obtain the estimate after smoothing the sparse cells. Poststratification via census information will match the estimate from the sample to the population. The combination of regression and poststratification is similar with the endogenous poststratification concept (Breidt, 2008; Dahlke et al., 2013). We introduce the MRP framework in detail.

3. Method

3.1. Multilevel regression and poststratification

In the basic setting, we are interested in estimating the population distribution of the survey outcome y. When the weighting process is transparent, we can directly include the weighting variable X into regression modeling for the survey outcome y. Here X is a q-dimensional vector of variables that affect the sampling design, nonresponse and coverage. Conditional on X, the distribution of inclusion indicator I is ignorable. Under MRP, the weighting variables X are discretized, and their cross-tabulation constructs the poststratification cells j, with population cell sizes N_j and sample cell sizes n_j, where J is the total number of poststratification cells (Little, 1991, 1993; Gelman and Little, 1997; Gelman and Carlin, 2001). Then the total population size is $N = \sum_{j=1}^{J} N_j$, and the sample size is $n = \sum_{j=1}^{J} n_j$.

Poststratification inference is different from design-base inference under stratified sampling by the fact that n_j's are now random functions of the sampling distribution I. In repeated sampling of I, there is a non-zero probability that $n_j = 0$ for some j. The usual resolution of this problem is to condition on n_j's observed in the realized sample, however, the sample inference is not design-unbiased conditionally on n_j's. The MRP framework assumes a model for n_j's and preserves design consistency.

The poststratification implicitly assumes that the units in each cell are included with equal probability. Suppose θ is the population estimand of interest, such as the overall or domain mean, and it can be expressed as a weighted sum over any subset or domain D of the poststrata,

$$\theta = \frac{\sum_{j \in D} N_j \theta_j}{\sum_{j \in D} N_j}, \tag{3.1}$$

where θ_j is the corresponding estimand in cell j. The proposed poststratified estimator will be of the general form,

$$\tilde{\theta}^{\mathrm{PS}} = \frac{\sum_{j \in D} N_j \tilde{\theta}_j}{\sum_{j \in D} N_j}, \tag{3.2}$$

where $\tilde{\theta}_j$ is the corresponding estimate in cell j. Various modeling approaches can be used to estimate the cell estimates, such as the flexible nonparametric Bayesian models and machine learning algorithms. Here, we illustrate using a hierarchical regression model.

In practice, survey weights are attached to each unit, even though they are not attributes of individual units. It is natural to generate unit-level weights based on the entire survey design, and use the weighted averages of the form, such as $\tilde{\theta} = \sum_{i=1}^{n} w_i y_i / \sum_{i=1}^{n} w_i$. Our goal here is to obtain an equivalent set of unit-level weights w_i through a model-based procedure for the estimation of $\tilde{\theta}^{\text{PS}}$ to connect weighting and poststratification. Therefore, regression models can be used to obtain $\tilde{\theta}_j$, poststratification accounts for the population information, and model-based weights are re-derived via Model (3.2).

In classical regression models, full poststratification is a special case, where the cell estimates are computed separately for each cell without any pooling effect, i.e., no pooling. For example, if we are interested in the population mean, then the cell means will be used as the cell estimates. Generally, classical regression models are conducted on cell characteristics without going to the extreme fitting separately for each cell. If more interactions among the characteristics are included, the resulted weights become more variable. On the other side, complete pooling ignores the heterogeneity among cells. Hierarchical regression models will smooth the variable estimates under partial pooling.

Gelman (2007) uses the exchangeable normal model as an illustration and shows that the poststratification estimate $\tilde{\theta}^{\text{PS}}$ for population mean can be expressed as a weighted average between the cell means and the global mean, which yields the unit weights, also as a weighted average between the completely smoothed weights, $w_j = 1$, and the weights from full poststratification, $w_j = (N_j/N)/(n_j/n)$. Hierarchical poststratification is approximately equivalent to shrinkage of weights through the shrinkage of the parameter estimates. The degree of shrinkage goes to zero as the sample increases, which implies that estimates from the model are design-consistent. However, further developments are necessary to handle large number of cells and deep interactions, and rigorously evaluate the performance of model-based weights.

In our application of the LSW study, the weighting variables include age (5 categories), ethnicity/race (5 categories), education (4 categories), sex (2 categories), poverty measure (5 categories), family size (4 categories), number of elder people (3 categories) and number of children (4 categories), in the family, and this results in $J = 5 \times 5 \times 4 \times 2 \times 5 \times 3 \times 4 \times 4 = 48,000$ poststrata. The majority of the poststratification cells will be empty or sparse due to limited sample size (2,002). The sample cell sizes are unbalanced. Often cells are arbitrarily collapsed or combined (Little, 1993) without theoretical justification. Recent model-based weighting smoothing procedures across cells could not handle such sparse cases (Elliott and Little, 2000). Xia and Elliott (2016) introduced a Laplace prior for weight smoothing across a modest number of poststrata based on inclusion probabilities but ignored the weighting variables and their hierarchy structure. Using the MRP framework, we account for the variable hierarchy structure to smooth and pool the estimates across the sparse and unbalanced cell sizes with a novel set of prior distributions.

3.2. Structured prior distribution

We introduce a structured prior distribution to improve MRP under the sparse and unbalanced cell structures, thus yielding stable model-based survey weights that account for design information. For now, suppose the population distribution of X is known, that is, we can obtain N_j's from the external data to describe a joint distribution of the weighting variables. Extension to unknown N_j's is discussed in Section 6. In practice, the number of poststratification cells J can be large, even

much larger than the sample size n. The weighting variables could affect the inclusion through a complex relationship or a differential response mechanism. Deep interactions are essential for complex relationship structure, but we cannot include all and have to select the predictive main effects and interactions.

Suppose the collected survey response is continuous, y_i, for $i = 1, \ldots, n$, and we are interested in the population mean $\bar{y}$ estimation. We use $(X^1, \ldots, X^J)^\top$ to represent the $J \times q$ predictor matrix in the population under the poststratification framework. For illustration, assume a normal distribution,

$$y_i \sim \mathrm{N}(\theta_{j[i]}, \sigma_y^2), \tag{3.3}$$

where $j[i]$ denotes the cell j that unit i belongs to. We can also consider unequal variances, allowing the cell scale σ_y to vary across cells, indexed as σ_j. For the prior specification of θ_j, one choice can be $\theta_j = X^j\beta$, and β is assigned with some prior distribution. In the hierarchical regression example of Gelman (2007), a multivariate normaldistribution is considered, $y \sim \mathrm{N}(X\beta, \Sigma_y)$ and $\beta \sim \mathrm{N}(0, \Sigma_\beta)$, where the covariates include all main effects and a few selected two-way interactions in X and the covariance matrix Σ_β is diagonal with different scales. However, the model is subject to misspecification, and the generated weights could be negative.

Since X^j consists different level indicators of the q discrete weighting variables, we can express the population cell mean θ_j as

$$\theta_j = \alpha_0 + \sum_{k \in S^{(1)}} \alpha_{j,k}^{(1)} + \sum_{k \in S^{(2)}} \alpha_{j,k}^{(2)} + \cdots + \sum_{k \in S^{(q)}} \alpha_{j,k}^{(q)}, \tag{3.4}$$

where $S^{(l)}$ is the set of all possible l-way interaction terms, and $\alpha_{j,k}^{(l)}$ represents the kth of the l-way interaction terms in the set $S^{(l)}$ for cell j. For example, $\alpha_{j,k}^{(1)}$'s with $k \in S^{(1)}$ refer to the main effects, $\alpha_{j,k}^{(2)}$'s with $k \in S^{(2)}$ being the two-way interaction terms, for cell j. This decomposition covers all possible interactions among the q weighting variables. When the cell structure is sparse, variable selection is necessary. In practical application, we recommend the initial inclusion of covariates and interactions with substantive importance and scientific interest in Model (3.4) and perform Bayesian variable selection under the proposed structured prior setting.

We induce structured prior distributions to be able to handle deep interactions and account for their hierarchy structure, where the high-order interaction terms will be excluded if one of the corresponding main effects is not selected. Larger main effects often lead to larger effects of the involved interaction terms. Ideally, more shrinkage should be put on the high-order interactions than that on the main effects, and the prior setting should reflect the nested structure. The challenge embodies the problem in Bayesian inference for group-level variance parameters in an ANOVA structure (Gelman, 2005, 2006). Volfovsky and Hoff (2014) introduce a class of hierarchical prior distributions for interaction arrays that can adapt to potential similarity between adjacent levels, where the covariance matrix for the high-order interactions is assumed as a Kronecker product of the covariance matrices of main effects after adjusting relative magnitudes. Our proposal extends by inducing more structure among the variance parameters, more shrinkage and smoothing effect to handle extremely large number of cells with unbalanced sizes than the generally balanced setting in Volfovsky and Hoff (2014), and improves the computation performance.

We start with independent prior distributions on the regression parameters α:

$$\alpha_{j,k}^{(l)} \sim \mathrm{N}(0, (\lambda_k^{(l)}\sigma)^2),$$

where $\lambda_k^{(l)}$ represents the local scale and σ is the global error scale, for $k \in S^{(l)}$ and $l = 1, \ldots, q$. The error scale is the same across the main effects and high-order interactions, while the local scales are different. The shrinkage effect is induced through the specification of local scales. We assume the local scale of high-order interactions is the product of those for the corresponding main effects after adjusting relative magnitudes.

$$\lambda_k^{(l)} = \delta^{(l)} \prod_{l_0 \in M^{(k)}} \lambda_{l_0}^{(1)},$$

where $\delta^{(l)}$ is the relative magnitude adjustment and $M^{(k)}$ is the collection of corresponding main effects that construct the kth l-way interaction in the set $S^{(l)}$. For example, the local scale of the three-way interaction among age, sex, and education, middle-aged men with college education, will be the product of those for the main effects on age, sex, and education, that is, the product of the local scale parameters for middle aged, men, and college educated, respectively.

We use the following hyperpriors on the scale parameters:

$$\begin{aligned}
\text{error scale: } & \sigma \sim \text{Cauchy}_+(0, 1) \\
\text{local scale for main effects: } & \lambda_k^{(1)} \sim \text{N}_+(0, 1) \\
\text{local scale for high-order interactions: } & \lambda_k^{(l)} = \delta^{(l)} \prod_{l_0 \in M^{(k)}} \lambda_{l_0}^{(1)} \\
\text{relative magnitude for high-order interactions: } & \delta^{(l)} \sim \text{N}_+(0, 1), \text{ for } l = 2, \ldots, q.
\end{aligned} \tag{3.5}$$

Here Cauchy_+ and N_+ denotes the positive part of the Cauchy and normal distributions, respectively. Gelman (2006) proposes the half-Cauchy prior for the scale parameter in hierarchical models, which has the appealing property that it allows scale values arbitrarily close to 0, with heavy tails allowing large values when supported by the data. When $\lambda_k^{(l)}$ is close to 0, the posterior samples of $\alpha_{j,k}^{(l)}$ are shrunk towards 0. The scale parameter for the high-order interaction terms will be 0 if any of the related scale parameters for the main effects is 0. The overall regularization effect is determined by the error scale and the multiplicative scale parameters of the corresponding main effects. We assign a noninformative prior distribution to the intercept term and weakly informative prior distributions to the two global error scale parameters (σ_y, σ), where $\sigma_y \sim \text{Cauchy}_+(0, 5)$.

Our proposed prior specification features the group selection of all possible level indicators for the same variable, similar to the group lasso (Yuan and Lin, 2006). We achieve the goal of variable selection under the similar specification with the Horseshoe prior distribution (Carvalho et al., 2010) and improve by setting up group selection and multiplicative scales for high-order interactions for sparsity gains. We introduce weakly informative half-Cauchy prior distributions to error scales and informative half-normal prior distributions to the local scale parameters, in the same spirit as in Piironen and Vehtari (2016), to improve parameter shrinkage estimation and computation efficiency. When the posterior estimation of the scale parameter is close to 0, indicating the variable is not predictive; post-processing can be done to exclude the variable from poststratification cell construction for dimension reduction. This class of priors allows for variable selection in high dimension and keeps the hierarchy structure among main effects and interactions.

3.3. Model-based weights

We can re-express the models (3.4) and (3.5) as the exchangeable normal model:

$$\theta_j \sim N(\alpha_0, \sigma_\theta^2), \quad \sigma_\theta^2 = \sum_{l=1}^{q} \sum_{k \in S^{(l)}} (\lambda_k^{(l)} \sigma)^2. \tag{3.6}$$

Conditional on the variance parameters, the posterior mean in the normal model with normal prior distribution is a linear function of data; thus we can determine *equivalent weights* w_i^*'s so that one can re-express the smoothed estimate $\sum_{j=1}^{J} N_j/N \tilde{\theta}_j$ as a classical weighted average, $\sum_{i=1}^{n} w_i^* y_i / \sum_{i=1}^{n} w_i^*$. Combining the posterior mean estimates for θ_j and the model-based estimate given in Model (3.2), Gelman (2007) derives the equivalent unit weights in cell j that can be used classically.

$$w_j \approx \frac{n_j/\sigma_y^2}{n_j/\sigma_y^2 + 1/\sigma_\theta^2} \cdot \frac{N_j/N}{n_j/n} + \frac{1/\sigma_\theta^2}{n_j/\sigma_y^2 + 1/\sigma_\theta^2} \cdot 1, \tag{3.7}$$

where the model-based weight is a weighted average between full poststratification weight without pooling (equal to $(N_j/N)/(n_j/n)$) and completely smoothing weight (equal to 1). The shrinkage effect is quantified as $1/(n_j\sigma_\theta^2/\sigma_y^2+1)$, where depends on the group and individual variances, as well as sample cell sizes. The model-based weights are random variables, and fully Bayesian inference will propagate the corresponding variability. We collect the posterior mean values and treat as the weights that can be used the same as classical weights.

3.4. Computation

The Bayesian hierarchical prediction and weighting inference procedure is reproducible and scalable. We implement the proposed structured prior distributions in the open source R package *rstanarm* (Goodrich and Gabry, 2017) released on CRAN. The computation codes are available online (Si et al., 2017) for public use. The fully Bayesian inference is realized via Stan. As open source and user friendly software, Stan contributes to the wide application of Bayesian modeling. Survey practitioners resist model-based approaches mainly due to computation burden. However, model-based methods are ready to face the new challenges on big survey data, such as unbalanced cell structure, combining multiple surveys and analyzing streaming data. The development of Stan can improve the generalization of model-based approach and provide the computational platform for the unified survey inference framework.

In our implementation, the Markov chain Monte Carlo samples mix well and the chains converge quickly. The fast computation speed widens the usability of model-based survey inference approaches. The proposed prior specification improves the stability for smoothed weights under partial pooling. To illustrate the capability of variable selection and hierarchy maintenance and the resulting efficiency gains, we compare the posterior estimation with that under independent prior setting but without the multiplicative scale constraint, which is similar with Horseshoe prior under group specification, called as independent prior distributions in the paper. We compare the model-based weights with classical weights in Section 4 and 5 to demonstrate the calibration for design-based properties. Furthermore, we illustrate the proposed improvement for domain estimation under unbalanced and sparse sample cell structure.

4. Simulation studies

We consider two main simulation scenarios: slightly unbalanced structure with a moderate number of poststratification cells and very unbalanced structure with a large number of poststratification cells. We evaluate the statistical validity of the model-based and weighted estimation for the finite population and domain inference to demonstrate the improved capability to solve the classical weighting problems. We consider model-based predictions under the structured prior (Str-P) and the independent prior (Ind-P) distributions. For weighted inference, we evaluate the estimation after applying the model-based weights under structured prior (Str-W) setting, model-based weights under independent prior (Ind-W) distributions, weights obtained via raking adjustment (Rake-W), classical poststratification weights (PS-W), and inverse probability of selection weighting (IP-W). We present the graphical diagnosis tools to compare the weights and weighted inference.

We borrow 2011 ACS survey of NYC adult residents treating it as the "population", and randomly draw samples out of it according to a pre-specified selection model without nonresponse. We collect covariates from ACS and simulate the outcome variable to obtain the true distribution as a benchmark. We implement the raking procedure by balancing the marginal distributions of the weighting variables in the selection model and generate the raking weights. The classical poststratification weights N_j/n_j's are obtained by matching the selected sample cell indices with those of the population cells. The selection model can provide the inverse probability of selection weights by matching the sampled unit indices. We also generate model-based weights under independent prior distributions for the main effects and high-order interaction terms of the ACS variables. The generated weights are normalized to average 1 for comparison convenience.

4.1. Slightly unbalanced structure

We first handle slightly unbalanced structure when the number of poststratification cells and the sample cell sizes are moderate. We implement repeated sampling process to investigate the frequentist properties of model-based predictions and weighted inferences. With little shrinkage effect on high-order interactions, the model-based prediction and weighting with structured prior distributions have similar performance with that under independent prior distributions, while outperforming the classical weighting approaches.

Assume three weighting variables are included in the selection and outcome models: age, ethnicity, and education. We discretize the three variables in ACS as *age* (18–34, 35–44, 45–54, 55–64, 65+), *eth* (non-Hispanic white, non-Hispanic black, Asian, Hispanic, other), and *edu* (less than high school, high school, some college, bachelor degree or above). The number of poststratification cells is $5 \times 5 \times 4 = 100$. We assume the outcome depends on deep interactions, including all the main effects, two-way and three-way interaction terms among the three weighting variables; and the selection indicator depends on the three main effects. The specific values of the coefficients are given in Tables A.2–A.3 in Appendix A. The values are set to reflect the strong correlations between the covariate and dependent variables. And the effects are not necessarily similar across the adjacent factor levels, different from the scenarios in Volfovsky and Hoff (2014). The error scale in the outcome model is set as 1, where the true value is always fully recovered from the posterior estimation. The data generation model is different from the estimation model, but the latter is flexible enough to cover the former since the dependency structure will be recovered by the estimation. The proposal is robust against model misspecification.

We repeat the sampling for 500 times. The sample sizes vary between 2141 and 2393 with

median 2288. Empty sample cells occur with spread-out selection probabilities (ranging from 0.001 to 0.269) over the repeated sampling process. The number of occupied cells in the sample is between 80 and 93 with median 87. The slightly unbalanced cell structure is common in practical surveys with simple and clean sampling design. The population quantities of interest include the overall mean, domain means across the $13\,(= 5 + 4 + 4)$ marginal levels of three weighting variables and domain mean for non-white youths (an example of interaction between age and ethnicity). We examine the absolute value of estimation bias, root mean squared error (RMSE), standard error (SE) approximated by the average value of standard deviations (Ave. SD) and nominal coverage rate of the 95% confidence intervals.

The outputs in Figure 4.1 show that the model predictions have the smallest RMSE, the smallest SE with reasonable coverage rates, and comparable bias among all the methods. All variables affecting the outcome and selection mechanism are included in the modeling to satisfy the Bayesian principle for ignorable sampling mechanism. The model will predict all the cell estimates including the empty cells in the sample, fully using the population information and poststratification cell structure. The weighting inference is conditional on the observed units within occupied cells, and thus less efficient than the model predictions. Generally, the model-based weighting inference has smaller RMSE and SE but more reasonable coverage rates than that with classical weighting. Raking adjustment is not valid for the domain estimation with large bias, large RMSE and poor coverage, even though the selection mechanism depends on only the main effects. The inverse probability of selection weighting inference tends to have large SE but low coverage rates, especially for domain estimation. The poststratification weighting inference is close to the model-based weighting estimation since the domain sizes are modestly large. The cell shrinkage effect towards no weighting is small (between 0 and 0.19 with mean 0.05) under slightly unbalanced design. The number of cases who are less than high school educated is small (around 80), resulting in large estimation bias and SE for the weighting inferences, but not in model-based predictions. The model-based predictions stabilize the small area estimation by smoothing.

Model prediction performs well and similarly under the structured prior distribution or independent prior distribution. This is expected due to the small shrinkage effect. The cell structure is slightly unbalanced, and the outcome and selection models depend on all the main effects and high-order interaction terms. But the structured prior setting yields more efficient inference than the independent prior setting with smaller SE. This improvement is obvious in the very unbalanced design as shown in the following simulation of Section 4.2.

Additionally, we considered nine cases with different survey outcome models and sample selection models depending on various predictors as in Table A.1 in Appendix A. The specific values of the coefficients are given in Tables A.2–A.3. The conclusions are consistent that the model-based prediction and weighting yield more efficient and precise inference than that under classical weighting, in particular for domain estimation.

4.2. Very unbalanced structure

Complex sampling design and response mechanisms tend to create very unbalanced data structures where most poststratification cells are sparse and empty. The proposed structured prior setting brings in strong regularization effect to stabilize the model prediction and improves the estimation efficiency, especially for domain estimation, outperforming the independent prior distributions. The posterior inference on scale parameters can inform variable selection to improve model interpretation. When the main effects are not predictive, neither are the related high-order interactions.

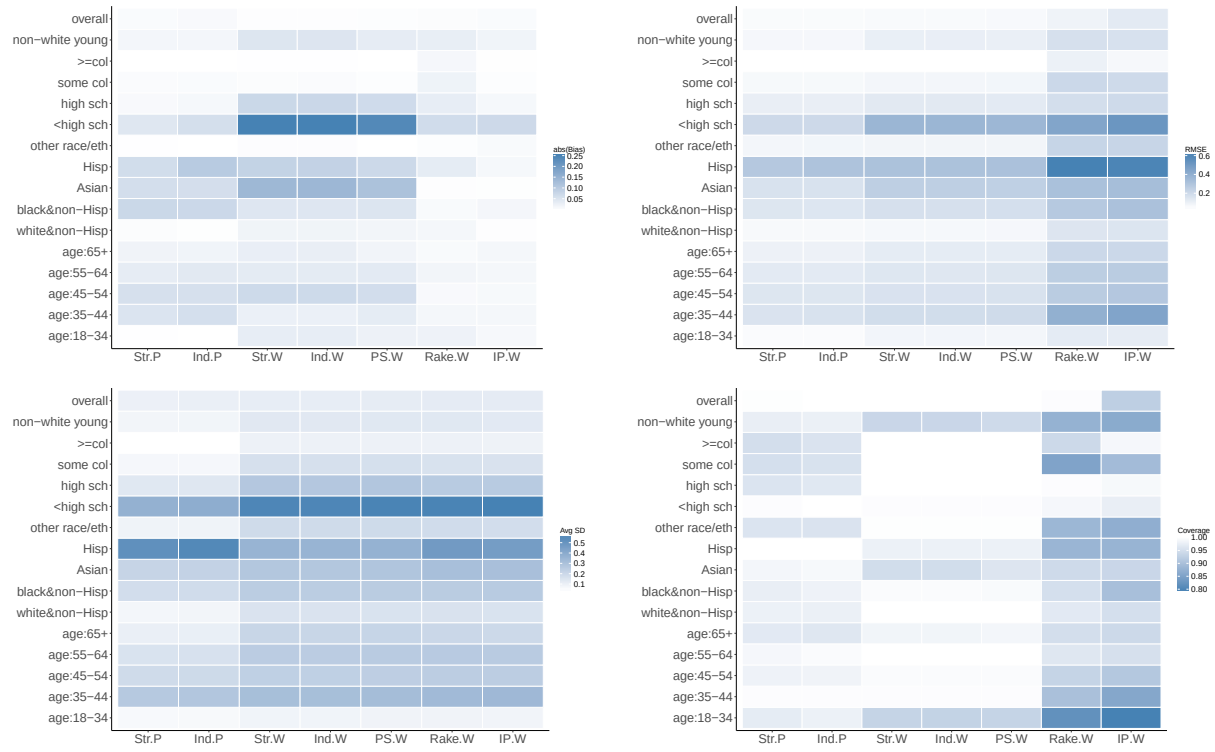

Figure 4.1: *Comparison of prediction and weighting performances on validity of finite population inference under slightly unbalanced design. The y-axis denotes different groups for the mean estimation. The x-axis includes two model-based prediction methods (Str-P, Ind-P), two model-based weighting methods (Str-W, Ind-W), and three classical weighting methods (PS-W, Rake-W, IP-W). Str-P: model-based prediction under the structured prior; Ind-P: model-based prediction under the independent prior distribution; Str-W: model-based weighting under structured prior; Ind-W: model-based weighting under independent prior distribution; Rake-W: weighting via raking adjustment; PS-W: poststratification weighting; and IP-W: inverse probability of selection weighting. The plots show that the model-based predictions outperform weighting with the smallest RMSE, the smallest SE, reasonable coverage rates, and comparable bias among all the methods. Model-based weighting inference has smaller RMSE and SE but more reasonable coverage rates than that with classical weighting.*

However, the posterior inference with independent prior distributions distorts the hierarchical structure between main effects and high-order interactions and hardly informs variable selection. The classical weighting inferences are highly variable in the sparse scenario.

Following the LSW, we collect eight weighting variables in the 2011 ACS-NYC data that affect sample inclusion: *age* (18–34, 35–44, 45–54, 55–64, 65+), *eth* (non-Hispanic white, non-Hispanic black, Asian, Hispanic, other), *edu* (less than high school, high school, some college, bachelor degree or above), *sex* (male, female), *pov* (one household income or poverty measure, poverty gap under 50%, 50–100%, 100–200%, 200–300%, more than 300%), *cld* (0, 1, 2, 3+ young children in the family), *eld* (0, 1, 2+ elders in the family), and *fam* (1, 2, 3, 4+ individuals in the family). The number of unique cells occupied by this classification is 8874, while the number of poststratification cells constructed by the full cross-tabulation is 48000.

In the simulation described in Table A.4 and Table A.5, the selection probability depends on the main effects of all variables, while the outcome depends on the main effects of five variables. The cell selection probabilities will be clustered, where some cells have the same selection probabilities. The error scale in the outcome model is set as 1. The selection probabilities fall between 0 and 0.90 with average 0.12, and we select 6374 units. Even though the sample sizes are large, the simulation creates a very unbalanced structure. The majority of the cells are empty, and 1096 of 1925 selected cells have one unit. Starting from an estimation model with sparsity, we assume the Model (3.4) for the cell estimations includes the main effects of the eight variables, eight two-way interactions and two three-way interactions. These terms are potentially important factors for weighting from the survey organizer's view. Our proposal can provide the insight of variable selection and then facilitate dimension reduction.

When only main effects are predictive, the posterior median values under the structured prior setting for the scales of the *cld*, *eld*, and *fam* are small (0.002, 0.003, 0.000), and the posterior median values for the scales of all high-order interactions are close to 0 (with magnitude smaller than or around 0.0001). The posterior mean of the error scale is 0.99 with SE 0.008, close to the true value 1. This is consistent with the simulation design. With independent prior distributions, however, the hierarchical structure between the main effects and high-order interaction terms is ignored. The posterior samples of scale parameters of the high-order interactions can be larger than that of the main effects. It is unclear about their predictive power and then hard to decide which terms to be selected. The posterior samples of the variance parameters under the independent prior distributions tend to be highly variable with heavy tails. For example, the variances of the main effects of age and sex have extremely large sampled values (14496 and 390000) and skewed distributions. For variables with a small number of levels, such as sex, the group-level variance estimation is sensitive to the prior distribution, and the independent prior distribution cannot regularize well. The structured prior distribution performs better by assuming the prior distributions share some common parameter and using more information for estimation, and then is able to stabilize the variance estimation. The structured prior setting yields more stable inference than the independent prior, and moreover can facilitate variable selection.

The proposed structured prior setting suggests that we exclude the non-predictive main effects and high-order interactions from the regression model for cell estimates, by either post-processing the posterior samples of the corresponding scales and coefficients to be 0 or refitting the updated model. In the simulation design, three variables affect the selection probability but are not related with the outcome. Inclusion of these variables into the regression model will increase the inference variability. The poststratification cell structure accounts for the eight variables to meet the

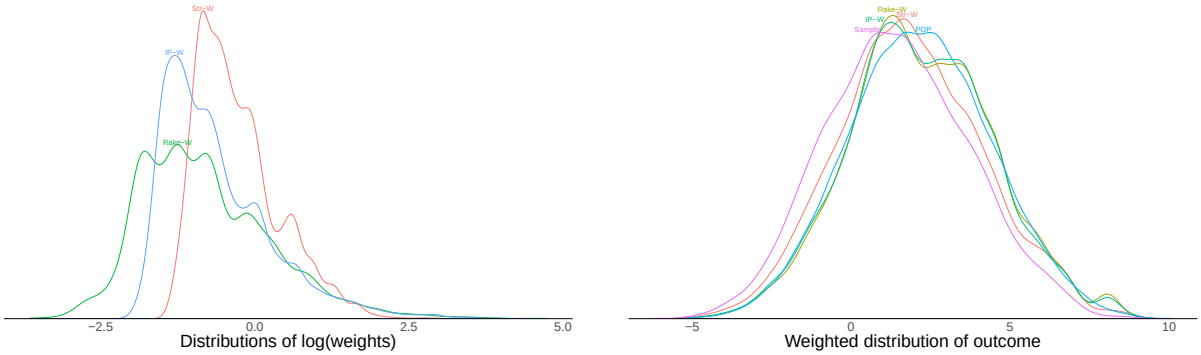

Figure 4.2: *Comparison of generated weights after logarithm transformation and weighted outcome distributions under a very unbalanced design. Str-W: model-based weighting under structured prior; Rake-W: weighting via raking adjustment; and IP-W: inverse probability of selection weighting. Sample: sample distribution of the outcome; and POP: population distribution of the outcome. The model-based weights are more stable and generate a more smoothed outcome distribution after weighting than the raking weights and the inverse probability of selection weights.*

ignorable sampling assumption. Further modification could be the exclusion of the three variables from the poststratification, which could make the assumption of ignorable sampling vulnerable but have efficiency gains. This is a tradeoff between efficiency and robustness that needs balance based on substantive interest. The selection of survey outcome variables in the weighting process needs further investigation, which we will elaborate in Session 6. We compared the inference with that after excluding the non-predictive terms and obtained similar outputs for the finite population and domain estimation, since the parameter estimates are close to 0 for the non-predictive terms. Here we present the outputs keeping such variables in the poststratification cell construction and the regression model.

First, we compare the generated weights by the model-based and classical methods. We collect the posterior samples of generated weights and present the posterior mean as the model-based weights. The model-based weights have smaller variability and narrower range than the classical weights, as shown in Figure 4.2. The iterative proportional fitting procedure does not coverage after the default 10 iterations that needs increasing. We examine the distribution of the outcome after accounting for the weights and compare with the population and sample distribution in the right plot of Figure 4.2. The sample distribution differs from the population distribution by underestimating the outcome values. The weighted distribution shifts towards the true population. The outcome distributions after weighting are similar among the model-based and classical methods, and the model-weights generate a smooth distribution of outcomes. This is reasonable as we expect the model-based weights perform similarly with classical weights on point estimation but improve efficiency by reducing the variability. The shrinkage effect under the structured prior distribution is large, between 0.86 and 1.00 with mean 0.90. The very unbalanced cell structure needs strong smoothing effect across cells. The model-based weights under the structured prior and independent distributions have similar distributions with the poststratification weights, so the latter two sets of weights are omitted in Figure 4.2.

We examine the inference for the overall mean and domain means across the marginal levels and for nonwhite young adults. The conclusions are the same as that in Section 4.1. Model-based

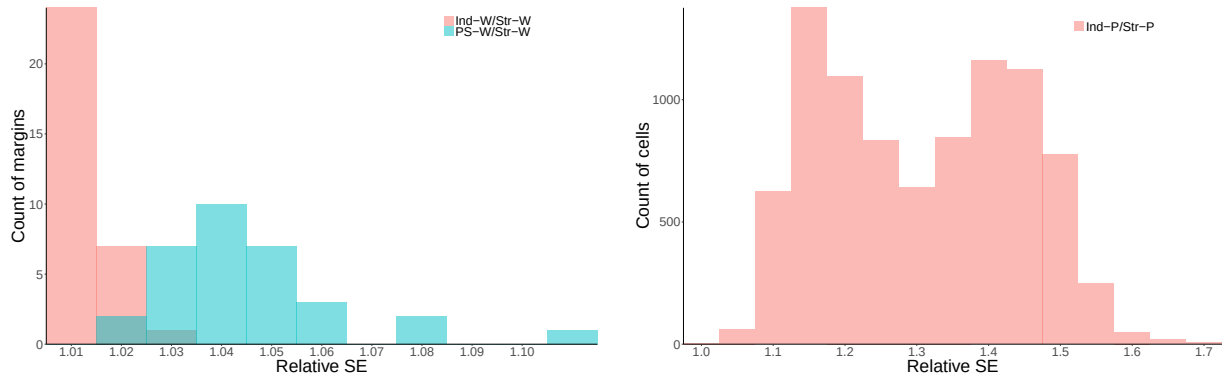

Figure 4.3: *Efficiency comparison of prediction and weighting performances on finite population domain inference under a very unbalanced design. The left plot examines the mean estimation across the margins defined by the eight weighting variables. The right plots presents the population cell mean estimation. The model-based weighting and prediction under the structured prior distribution yield smaller SE than those under independent prior. Model-based weighting yields smaller SE than poststratification weighting.*

prediction outperforms weighting inference with smallest bias and SE. The benefit can be explained by that the model uses the population information for empty cell prediction under regularization. Model-based weighting inference has smaller SE than that with classical weighting. Even when the selection probabilities depend on only main effects, raking yields small bias but performs badly with large SE.

Under the very unbalanced design, the model-based weighting inference under structured prior setting is more efficient than that under independent prior setting or with poststratification weights. We compare the SE of the marginal mean estimates of the eight weighting variables from the three weighting methods and plot the relative ratios in the left plot of Figure 4.3. The model-based weighting inference has smaller SE than the poststratification weighting, and the weighting under structured prior setting has the smallest SE. Because the sample sizes and the domain sizes are large and the data generation model is sparse, the model-based weighting inference has a little but not much improvement over the poststratification weighting inference due to small smoothing effect.

The model-based prediction and inference under the structured prior setting are more efficient than that under the independent prior setting. The SEs are smaller with the structured prior than those with the independent prior in the right plot of Figure 4.3. To demonstrate the efficiency gain, we look at the SEs for the population cell estimates. The Bayesian structural inference generally has smaller variability than that with independent prior, especially in the sparse scenarios.

We assume different outcome and selection models with different covariates with scenarios summarized in Table A.4 and achieve the same evaluation conclusions.

5. Application to Longitudinal Study of Wellbeing

With the background introduced in Section 2, we apply the prediction and weighting inference to the NYC Longitudinal Study of Wellbeing. We match the LSW to the adult population via the ACS. We would like to conduct finite population and domain inference and generate weights

allowing for general analysis use. The outcome of interest is the self-reported score of life satisfaction on a 1–10 scale. We model the outcome as normally distributed, which is not quite correct given that the responses are discrete, but should be fine in practice for the goal of estimating averages. We first include the same eight weighting variables to construct the poststratification cells and use the same estimation model as those in Section 4.2 under the structured prior setting. The posterior inference shows that the variables *sex, cldx, eldx*, and *psx* are not predictive, and neither are the related high-order interactions. The scale estimates of such terms have posterior median values close to 0 and several large values as long tails. The posterior samples of scales for several high-order interactions among the remaining four variables concentrate around 0, showing these quantities are not predictive. Another complexity is that, for the sample cells of the LSW, the corresponding population cells are not available in the ACS data. This could happen because the sampling framework is not the ACS survey. The population information is unknown for such cells, and untestable assumptions have to be made. The model fitting improves after variable selection when we check the prediction errors for cell estimates.

Hence, we use four weighting variables after selection, *age, eth, edu* and *pov*, which constructs 500 poststratification cells. The 2002 units in the LSW spread out in 359 cells. The largest sample cell has 86 units, while 92 cells have only one unit. The covariates in the model (3.4) for cell estimates include the main effects of the four variables, five two-way interactions (*age * eth*, *age * edu*, *eth * edu*, *age * inc* and *eth * inc*), and two three-way interactions (*age * eth * edu* and *age * eth * inc*). We implement the fully Bayesian inference with the structured prior distributions. We are interested in estimating the average score of life satisfaction for overall and several subgroups of NYC adults, and construct weights for general analysis purposes using the LSW.

The posterior median of the unit scale inside cells σ_y is 1.93 with 95% credible interval $[1.87, 1.99]$. The posterior median of the group variance σ_θ^2 is 0.63 with 95% credible interval $[0.42, 1.04]$. These lead to moderately large shrinkage effects between 0.11 and 0.90 with mean 0.30 across cells. The moderate shrinkage effect makes sense based on the four weighting variables and up to three-way interactions being included. The posterior mean values of the model-based weights are presented in the left plot of Figure 5.1. We can generate the raking weights after adjustment for the marginal distributions of the four weighting variables and poststratification weights based on the ACS data. The population information is obtained after applying the ACS personal weights.

Comparing with the classical weights, our model-based weights have smaller variability with standard deviation 0.32 and the ratio of the maximum and minimum value 3.87, and these values are much smaller than those for the raking and poststratification weights, as shown in Table 5.1. The right plot in Figure 5.1 shows the distribution of the lift satisfaction score after weighting. The model-based weighted distributions and classically weighted distributions are similar as expected, which is consistent with the results in Section 4.2. The weighting process adjusts for the sample distribution by upweighting the high scores and downweighting the low scores. The LSW oversamples poor residents who tend not be satisfied with life, and the weighting adjustment balances the discrepancy.

Table 5.1 and Figure 5.2 present the finite population and domain inference. The average score of life satisfaction for NYC adults is 7.24 with standard error 0.05, predicted by the structural model. The estimate is similar with that under model-based weighting and raking inferences, but lower than the poststratification weighting inference. However, the difference is not significant. For example, the structural model predicts the average score of life satisfaction for middle-aged, college-educated whites with income more than three times the poverty level as 7.40 with standard

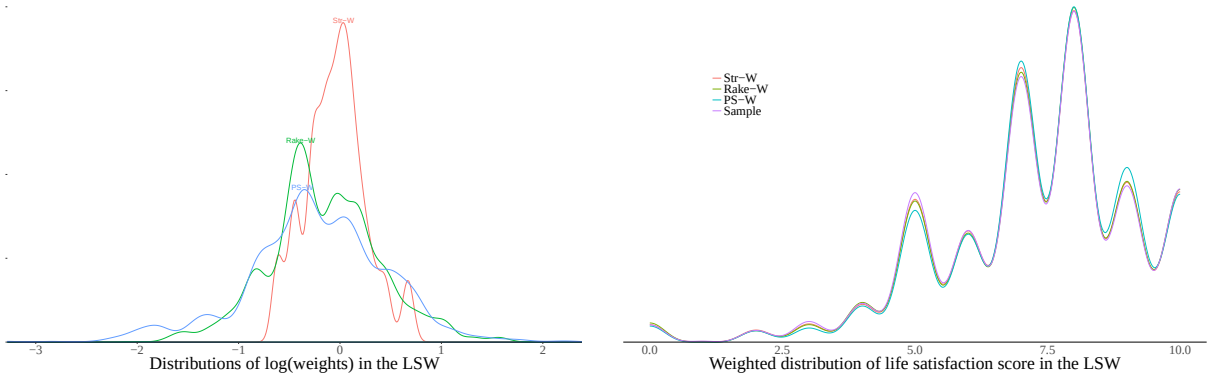

Figure 5.1: *Comparison of generated weights after logarithm transformation and weighted distributions of life satisfaction score in the LSW. Str-W: model-based weighting under structured prior; Rake-W: weighting via raking adjustment; IP-W: inverse probability of selection weighting, and Sample: sample distribution of the outcome. The weighted distributions are similar between model-based weights and classical weights, but model-based weights are more stable than classical weights.*

Table 5.1: *Comparison of prediction and weighting performances on estimating various domain averages for life satisfaction in the LSW. Str-P: model-based prediction under the structured prior; Str-W: model-based weighting under structured prior; Rake-W: weighting via raking adjustment; and PS-W: poststratification weighting.*

	Str-P	Str-W	Rake-W	PS-W
SD of weights / mean of weights		0.32	0.66	0.80
Max weight / min weight		3.87	81.28	274.65
Overall average for NYC adults ($n = 2002$)				
Est	7.24	7.23	7.24	7.30
SE	0.05	0.05	0.05	0.06
Average for middle-aged, college-educated whites with poverty gap $> 300\%$ ($n = 222$)				
Est	7.40	7.34	7.34	7.34
SE	0.10	0.11	0.11	0.11
Average for elders with poverty gap $< 200\%$ ($n = 154$)				
Est	7.37	7.52	7.49	7.53
SE	0.15	0.18	0.19	0.22
Averages for blacks with poverty gap $< 50\%$ ($n = 57$)				
Est	7.01	7.16	7.30	7.16
SE	0.18	0.26	0.28	0.29

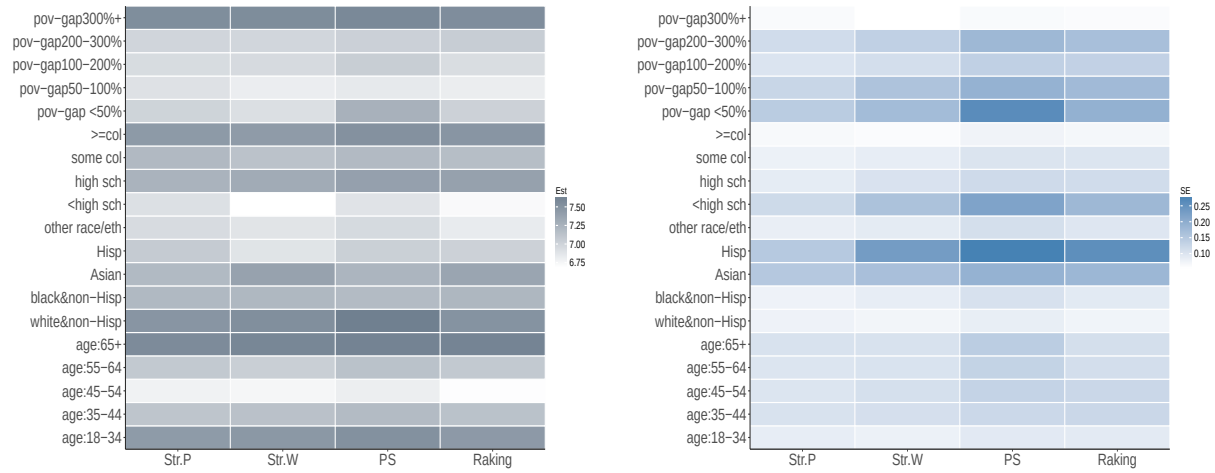

Figure 5.2: *Comparison of predictions and weighting performances on estimating life satisfaction score across the margins of four weighting variables in the LSW. Str-P: model-based prediction under the structured prior; Str-W: model-based weighting under structured prior; Rake-W: weighting via raking adjustment; and PS-W: poststratification weighting. Model-based predictions and weighting generate different estimates for several subsets and are generally more efficient comparing with classical weighting.*

error 0.10, higher than that under weighting inferences. Nevertheless, the predicted scores for elder with relatively low income (7.37 with SE 0.15) and low-income black New Yorkers (7.01 with SE 0.18) are lower than those under weighting inferences. The discrepancy could be explained by the nonrepresentativeness of the LSW and the deep interactions included by the model. The subgroup of individuals who are middle-aged, college-educated whites may be undercovered in the LSW—as empty poststratification cells occurring—with overcoverage among elderly poor blacks. Weighting the collected samples cannot infer or extrapolate inference on those who are not present in the survey. Though the differences are not significant, inferences conditioning on the collected samples are not design-consistent, especially for the empty cell estimates. Figure 5.2 shows the model-based prediction yields higher score for young, highly educated and Hispanic NYC adults, but lower score for those with poverty gap $< 50\%$, comparing with the weighted inference.

The SEs are similar for the overall mean estimation between predictions and various weighting inferences because of the large sample size. For domain estimation, the model-based prediction and weighting are more efficient than that with raking and poststratification weighting, and the model-based prediction has the smallest standard error. The efficiency gains of model-based prediction and weighting are further demonstrated by domain mean estimation for life satisfaction scores across the marginal levels of four weighting variables, shown in Figure 5.2. The model-based prediction and weighting particularly improve small domain estimation and increase the efficiency.

Survey practitioners often compare the weighted distribution of socio-demographics with the population distribution to check the weighting. While weighting diagnostics need further research and management, we follow this routine to compare the model-based and classical weights. We calculate the Euclidean distances between the weighted distributions and the population distribution for the main effects and high-order interactions among the four weighting variables in the LSW, shown in Table A.6 in Appendix A. The weighted distributions are generally close to the true

distributions. Raking focuses on adjusting for the marginal distributions of weighting variables but distorts the joint distributions, where the dependency structure is determined only by the sample without calibration. The poststratification weighting adjusts for the joint distribution, but empty cells in the sample present from the exact matching. The unbalanced cell structure yields unstable inference. The model-based weighting smooths the poststratification weightings and outperforms raking to match the distributions of three-way and four-way interaction terms. Practitioners often rely the marginal distributions to evaluate weighting performances, thus in favor of raking. However, raking yields high variable and potentially biased inferences, shown in the Section 4, even in the cases when raking adjustment is correct. Modification of model-based weighting to satisfy such desire on matching marginal distributions will be a future extension to incorporate constrains.

6. Discussion

We combine Bayesian prediction and weighting as a unified approach to survey inference. Multilevel regression with structured prior distributions and poststratification on the population inference yield efficient and design-consistent estimation. The computation is implemented via Stan and disseminated through the R package *rstanarm* for public use, and the software development promotes the model-based approaches in survey research and operational practice. We construct stable and calibrated model-based weights to solve the problems of classical weights. This article builds up the model-based prediction and weighting framework and serves as the first contribution to evaluate the statistical properties of model-based weights and compare the performances with classical weighting. Model-based weights are smoothed across poststratification cells and improve small domain estimation.

The structured prior uses the hierarchical structure between main effects and high-order interaction terms to introduce multiplicative constraints on the corresponding scale parameters and informs variable selection. Model improvement can be done after post-processing the posterior inferences. The Bayesian structural model yields more stable inference than that with independent prior distributions. Furthermore, the unified prediction and weighting approach is well equipped to deal with complex survey designs and big data in surveys, such as streaming data and combining multiple survey studies.

The general MRP framework is open to flexible modeling strategies. In this article, we illustrate by a regression model with all variables of interest and the high-order interactions and incorporate structured prior distributions for regularization. Other approaches, such as nonparametric models and machine learning tools, can be implemented under the MRP framework, being robust against model misspecification. Si et al. (2015) use Gaussian process regression models to borrow information across poststratification cells based on the distances between the inverse inclusion probability weights. Further extensions include applying such flexible approaches to weight smoothing and deriving the model-based weights.

The broad application opportunities come with various challenges that need further investigation. The model-based weights are outcome dependent, which improves the efficiency but potentially reduces the robustness. Survey organizers prefer a set of weights than can be used for general analysis purpose, without being sensitive to outcome selection. We can compare different weights constructed by several important outcomes and conduct sensitivity analysis. When the model-based weights give different inference conclusions, we recommend choosing the set of weights that generate the most reasonable results, with scientific reasoning and be consistent with the

Table A.1: *Covariates in the outcome (O) and selection (S) models for slightly unbalanced design.*

	Case 1		Case 2		Case 3		Case 4		Case 5		Case 6		Case 7	
	O	S	O	S	O	S	O	S	O	S	O	S	O	S
age	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
eth	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓
edu	✓	✓	✓	✓	✓		✓	✓	✓		✓	✓	✓	
age*eth	✓			✓	✓	✓				✓				✓
age*edu	✓			✓	✓						✓		✓	
eth*edu	✓			✓	✓									
age*eth*edu	✓			✓	✓									

population inference.

The weighted marginal distributions of the weighting variables are a bit different from the population inferences, as in Section 5, which does not meet the usual weighting diagnosis standard of survey organizers. The model-weights tend to match the joint distribution of the weighting variables to that in the population, but weight smoothing may bring in bias. Tradeoff constraints can be induced to the model to match the marginal distributions.

Another practical challenge is that, the population distribution of the weighting variables may be unknown, that is, the population poststratification cell sizes N_j's are unknown. A supplemental model is needed to allow estimation of this information from the sample. When marginal distributions are available, Little and Wu (1991) discuss an equivalent model approach for raking. Auxiliary variables or additional information on the population can be included in the Bayesian framework. Some auxiliary variables' population distribution may not be available in the census database, such as the number of phones, and we can estimate from other surveys as reference samples.

The model-based predictions and weighting inferences need further extensions to handle non-continuous outcome, inference on regression coefficients and non-probability or informative sampling designs (Kim and Skinner, 2013). It will be useful to link these ideas on survey inference with biostatistical and econometric literatures on inverse propensity score and doubly robust weighting (Kang and Schafter, 2007).

Acknowledgements

We thank the U.S. National Science Foundation and the Office of Naval Research for grant support.

A. Simulation designs

Here we present the simulation designs, coefficient values, and comparison on the weighted distributions of socio-demographics as a supplement to Section 4 and Section 5.

References

ACS Weighting Method (2014). *American Community Survey Design and Methodology, Chapter 11: Weighting and Estimation*. United States Census Bureau.

Table A.2: *Assumed regression coefficients in the* outcome *model for the simulation using a slightly unbalanced design.*

	All	Main effects	Two variables
age	(0.5 1.375 2.25 3.125 4)	(0.5 1.375 2.25 3.125 4)	(0.5 1.375 2.25 3.125 4)
eth	(-2 -1 0 1 2)	(2 -1 0 1 2)	$\vec{0}$
edu	(3 2 1 0)	(3 2 1 0)	(3 2 1 0)
age*eth	(4 2 1 1 3 3 2 1 1 1 2 3 2 2 1 4 4 3 2 3 2 4 1 4 1)	$\vec{0}$	$\vec{0}$
age*edu	(-2 -1 2 2 1 -2 2 1 0 -2 1 -2 -1 2 1 -1 -1 2 0 2)	$\vec{0}$	(2 0 -2 -2 1 1 -1 -2 -2 -1 -1 1 0 -1 -1 2 2 1 -1 0)
eth*edu	(1 -2 0 -3 -1 0 -1 -2 0 -1 -3 -3 0 -1 -1 0 0 -1 0 -1)	$\vec{0}$	$\vec{0}$
age*eth*edu	(-1 -0.5 0.5 -1 -1 -0.5 -1 0 -1 0 -1 0 1 1 0.5 1 1 -1 -1 0 -1 -0.5 -0.5 -1 1 -1 -0.5 -1 1 0 0.5 0.5 1 0.5 1 1 1 0.5 1 0 0 -0.5 0 1 -1 -1 0 -1 -1 -1 -0.5 -0.5 0 1 -1 0 0 -0.5 1 -0.5 0.5 -1 1 0 1 0 -1 0 -0.5 1 -0.5 -1 -0.5 0 0.5 -0.5 1 0.5 -0.5 0.5 0 1 0 1 0.5 0.5 0.5 0 0 -0.5 1 -1 0 1 1 1 1 -0.5 -1 -1)	$\vec{0}$	$\vec{0}$

Table A.3: *Assumed regression coefficients in the* selection *model for the simulation using a slightly unbalanced design.*

	All	Main effects	Two variables
Intercept	-2	-2	-2
age	(-2 -1.75 -1.5 -1.25 -1)	(0 0.5 1 1.5 2)	(-2 -1.5 -1 -0.5 0)
eth	(-1 -0.25 0.5 1.25 2)	(-2 -1.5 -1 -0.5 0)	(-1 -0.5 0 0.5 1)
edu	(0 0.67 1.33 2)	(0 1 2 3)	$\vec{0}$
age×eth	(1 1 -1 1 -1 1 -1 0 0 -1 0 0 -1 1 0 0 -1 1 1 -1 -1 0 1 -1 1)	$\vec{0}$	(-1 1 1 1 -1 -1 -1 0 -1 -1 -1 -1 1 -1 -1 0 1 1 -1 1 -1 -1 1 0 0)
age×edu	(0 1 -1 -1 0 1 1 0 1 0 1 -1 -1 1 1 -1 0 -1 1 1)	$\vec{0}$	$\vec{0}$
eth×edu	(-1 -1 0 -1 -1 1 1 1 1 0 -1 0 -1 0 -1 1 0 -1 -1 -1)	$\vec{0}$	$\vec{0}$
age×eth×edu	(0.8 -0.4 0.6 -0.2 0.8 0.2 0.4 0.8 0.4 -0.6 -0.8 -0.4 -0.8 -0.4 0.4 -1 0.6 -0.8 -0.6 0.6 -0.2 0.2 0.6 -0.6 0 0 -1 -0.2 0.6 0.8 -0.4 0.2 -0.8 0.4 0.6 -0.6 0.8 0 0.2 -1 1 0.4 0 0.8 -0.2 0 0 0.6 -0.8 -0.8 -0.2 0.4 -1 -0.8 1 -0.2 0 0.8 0.6 0.8 -0.2 -0.2 -0.8 1 0.8 0.8 -0.4 -0.8 0.4 -0.4 1 -0.6 -1 -0.6 -0.2 1 1 -0.2 1 0.6 0.4 0.8 0.2 -0.2 -0.6 0 0.8 -0.4 0.4 0.4 0.6 -1 -0.8 -0.8 1 1 0.4 0.6 0.4 0.8)	$\vec{0}$	$\vec{0}$

Table A.4: *Covariates in the outcome (O) and selection (S) models for a very unbalanced design.*

	Case 1		Case 2		Case 3		Case 4	
	O	S	O	S	O	S	O	S
age	✓	✓	✓	✓	✓	✓	✓	✓
eth	✓	✓	✓	✓	✓	✓	✓	✓
edu	✓	✓	✓	✓	✓	✓	✓	✓
sex	✓	✓	✓	✓	✓	✓	✓	✓
pov	✓	✓	✓	✓	✓	✓	✓	✓
cld		✓		✓		✓	✓	✓
eld	✓	✓		✓	✓	✓	✓	✓
fam	✓	✓		✓	✓	✓	✓	✓
age*eth	✓	✓			✓			✓
age*edu	✓	✓			✓			✓
eth*edu	✓	✓			✓			✓
eth*pov	✓	✓			✓			✓
age*pov	✓	✓			✓			✓
pov*fam	✓	✓			✓			✓
pov*eld	✓	✓			✓			✓
pov*cld		✓						✓
age*eth*edu	✓	✓			✓			✓
age*eth*pov	✓	✓			✓			✓

Table A.5: *Assumed regression coefficient values for the outcome (O) and selection (S) models for a very unbalanced design.*

	O	S
age	(2 0 -2 -2 1)	(0 0.75 1.5 2.25 3)
eth	(1 -1 -2 -2 -1)	(-1 -0.5 0 0.5 1)
edu	(-1 1 0 -1)	(0 0.67 1.33 2)
sex	(-1 2)	(-1 0)
pov	(2 1 -1 0 -1)	(0 1 2 3 4)
cld	$\vec{0}$	(-1 -0.33 0.33 1)
eld	$\vec{0}$	(-2 -1 0)
fam	$\vec{0}$	(-1 -0.67 -0.33 0)

Table A.6: *Euclidean distances between the weighted distributions and the population distribution. Str-W: model-based weighting under structured prior; Rake-W: weighting via raking adjustment; and PS-W: poststratification weighting.*

	Str-W	PS-W	Rake-W
age	0.04	0.02	0.00
eth	0.08	0.06	0.00
edu	0.08	0.03	0.00
inc	0.02	0.02	0.00
*age * eth*	0.05	0.03	0.05
*age * edu*	0.05	0.02	0.05
*age * inc*	0.03	0.01	0.03
*eth * edu*	0.06	0.04	0.05
*eth * inc*	0.04	0.04	0.03
*edu * inc*	0.06	0.03	0.04
*age * eth * edu*	0.03	0.02	0.05
*age * eth * inc*	0.03	0.02	0.04
*age * edu * inc*	0.03	0.01	0.04
*eth * edu * inc*	0.04	0.02	0.04
*age * eth * edu * inc*	0.02	0.01	0.04

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

Breidt, F. J. (2008). Endogenous post-stratification in surveys: Classifying with a sample-fitted model. *Annals of Statistics 36*, 403–427.

Breidt, F. J., G. Claeskens, and J. D. Opsomer (2005). Model-assisted estimation for complex surveys using penalized splines. *Biometrika 92*, 831–846.

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

Chambers, R. L., A. H. Dorfman, and T. E. Wehrly (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association 88*, 260–269.

Chen, Q., M. R. Elliott, D. Haziza, Y. Yang, M. Ghosh, R. Little, J. Sefransk, and M. Thompson (2017). Weights and estimation of a survey population mean: A review. *Statistical Science 32*(2), 227–248.

Dahlke, M., F. Breidt, J. Opsomer, and I. V. Keilegom (2013). Nonparametric endogenous post-stratification in surveys. *Statistica Sinica 23*, 189–211.

Deville, J. C. and C. E. Särndal (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association 87*(418), 376–382.

Deville, J. C., C. E. Sarndal, and O. Sautory (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association 88*(423), 1013–1020.

Elliott, M. R. (2007). Bayesian weight trimming for generalized linear regression models. *Journal of Official Statistics 33*(1), 23–34.

Elliott, M. R. and R. J. Little (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics 16*(3), 191–209.

Firth, D. and K. E. Bennett (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society Series B 60*, 3–21.

Fuller, W. (2009). *Sampling Statistics*. Wiley.

Gelman, A. (2005). Analysis of variance: why it is more important than ever (with discusion). *Annals of Statistics 33*(1), 1–53.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis 3*, 515–533.

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

Gelman, A. and J. B. Carlin (2001). Poststratification and weighting adjustments. In R. Groves, D. Dillman, J. Eltinge, and R. Little (Eds.), *Survey Nonresponse*.

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

Gelman, A. and T. C. Little (1998). Improving on probability weighting for household size. *Public Opinion Quarterly 62*, 398–404.

Ghitza, Y. and A. Gelman (2013). Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Sicience 57*(3), 762–776.

Ghosh, M. and G. Meeden (1997). *Bayesian Methods for Finite Population Sampling*. CRC Press.

Goodrich, B. and J. S. Gabry (2017). rstanarm: Bayesian applied regression modeling via Stan. https://cran.r-project.org/web/packages/rstanarm/.

Groves, R. and M. Couper (1995). Theoretical motivation for post-survey nonresponse adjustment in household surveys. *Journal of Official Statistics 11*, 93–106.

Henry, K. and R. Valliant (2012). Comparing alternative weight adjustment methods. In *Proceedings of the Section on Survey Research Methods*. American Statistical Association.

Hoffman, M. D. and A. Gelman (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research 15*, 1351–1381.

Holt, D. and T. M. F. Smith (1979). Post stratification. *Journal of the Royal Statistical Society Series A 142*(1), 33–46.

Kang, J. D. Y. and J. L. Schafter (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science 22*(4), 523–539.

Kim, J. K. and C. J. Skinner (2013). Weighting in survey analysis under informative sampling. *Biometrika 100*(2), 385–398.

Kott, P. (2009). Calibration weighting: combining probability samples and linear prediction models. In D. Pfeffermann and C. R. Rao (Eds.), *Handbook of Statistics, Sample Surveys: Design, Methods and Application*, Volume 29B. Elsevier.

Little, R. (1983). Comment on "An evaluation of model-dependent and probability-sampling inferences in sample surveys", by M. H. Hansen, W. G. Madow and B. J. Tepping. *Journal of the American Statistical Association 78*, 797–799.

Little, R. (1991). Inference with survey weights. *Journal of Official Statistics 7*, 405–424.

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

Little, R. (2004). To model or not to model? Competing modes of inference for finite population sampling inference for finite population sampling. *Journal of the American Statistical Association 99*, 546–556.

Little, R. (2011). Calibrated Bayes, for statistics in general, and missing data in particular. *Statistical Science 26*(2), 162–174.

Little, R. and M. Wu (1991). Models for contingency tables with known margins when target and sampled populations differ. *Journal of the American Statistical Association 86*, 87–95.

Park, D. K., A. Gelman, and J. Bafumi (2005). State-level opinions from national surveys: Poststratification using multilevel logistic regression. In J. E. Cohen (Ed.), *Public Opinion in State Politics*. Stanford University Press.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review 61*(2), 317–337.

Piironen, J. and A. Vehtari (2016). On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. https://arxiv.org/abs/1610.05559.

Potter, F. A. (1988). Survey of procedures to control extreme sample weights. In *Proceedings of the Section on Survey Research Methods*, pp. 453–458. American Statistical Association.

Potter, F. A. (1990). A study of procedures to identify and trim extreme sampling weights. In *Proceedings of the Survey Research Methods Section*, pp. 225–230. American Statistical Association.

Rubin, D. B. (1976). Inference and missing data (with discussion). *Biometrika 63*, 581–592.

Rubin, D. B. (1983). Comment on "An evaluation of model-dependent and probability-sampling inferences in sample surveys," by M. H. Hansen, W. G. Madow and B. J. Tepping. *Journal of the American Statistical Association 78*, 803–805.

Si, Y. and A. Gelman (2014). Survey weighting for New York longitudinal survey on poverty measure. Technical report, Columbia University.

Si, Y., N. S. Pillai, and A. Gelman (2015). Nonparametric Bayesian weighted sampling inference. *Bayesian Analysis 10*(3), 605–625.

Si, Y., R. Trangucci, and J. S. Gabry (2017). Computation codes for manuscript "Bayesian hierarchical weighting adjustment and survey inference". https://github.com/yajuansi-sophie/weighting.

Stan Development Team (2017a). Stan: A C++ library for probability and sampling. http://mc-stan.org.

Stan Development Team (2017b). Stan modeling language user's guide and reference manual. http://mc-stan.org.

Valliant, R., A. Dorfman, and R. Royall (2000). *Finite Population Sampling and Inference*. Wiley.

Volfovsky, A. and P. Hoff (2014). Hierarchical array priors for ANOVA decompositions of cross-classified data. *Annals of Applied Statistics 8*(1), 19–47.

Wimer, C., I. Garfinkel, M. Gelblum, N. Lasala, S. Phillips, Y. Si, J. Teitler, and J. Waldfogel (2014). Poverty tracker—monitoring poverty and well-being in NYC. Columbia Population Research Center and Robin Hood Foundation.

Xia, X. and M. R. Elliott (2016). Weight smoothing for generalized linear models using a Laplace prior. *Journal of Official Statistics 32*(2), 507–539.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B 68*(1), 49–67.