Bayesian inference for the finite population total from a heteroscedastic probability proportional to size sample

Sahar Z Zangeneh* Roderick J.A. Little†

Abstract

We study Bayesian inference for the population total in probability-proportional-to-size (PPS) sampling. The sizes of non-sampled units are not required for the usual Horvitz-Thompson or Hajek estimates, and this information is rarely included in public use data files. Zheng and Little (2003) showed that including the non-sampled sizes as predictors in a spline model can result in improved point estimates of the finite population total. In Little and Zheng (2007), the spline model is combined with a Bayesian bootstrap (BB) model for the sizes, for point estimation when the sizes are only known for the sampled units. We further develop their methods by (a) including an unknown parameter to model heteroscedastic error variance in the spline model, an important modeling feature in the PPS setting; and (b) developing an improved Bayesian method for including summary information about the aggregate size of non-sampled units. Simulation studies suggest that the resulting Bayesian method, which includes information on the number and total size of the non-sampled units, recovers most of the information in the individual sizes of the non-sampled units, and provides significant gains over the traditional Horvitz-Thompson estimator. The method is applied on a data set from the US Census Bureau.

Key Words: Bayesian bootstrap; Heteroscedasticity; Penalized spline; Probability proportional to size; Metropolis-Hastings within Gibbs

1. Introduction

1.1 Background

We consider inference for probability proportional to size (PPS) sampling, where units from a finite population are sampled with probabilities proportional to a size variable X. Let y_i, $i = 1, \ldots, N$ be the survey (or outcome) variable of the ith unit, where $N < \infty$ is the number of units in the population and let I_i, $i = 1, \ldots, N$ be the inclusion indicator variable of the ith unit. We consider inference about the finite population total $Q(Y) = \sum_{i=1}^{N} y_i$, where $Y = (y_1, \ldots, y_N)$.

In the design-based or randomization approach (Cochran, 2009), inferences are based on the distribution of $I = (I_1, \ldots, I_N)$, and the outcome variables $y_1, \ldots, y_N$ are treated as fixed quantities. This is the traditional approach in the survey literature and is desirable for its lack of reliance on distributional assumptions. It automatically takes features of the survey design into account and yields reliable inferences for large samples; however, this approach is generally asymptotic and can be inefficient in small samples.

*Department of Statistics, The University of Michigan, Ann Arbor, MI, U.S.A.

†Department of Biostatistics, The University of Michigan, Ann Arbor, MI, U.S.A.

On the other hand, the model-based approach treats both $I = (I_1, \ldots, I_N)$ and $Y = (y_1, \ldots, y_N)$ as random variables. A model is assumed for the survey outcomes Y with underlying parameters θ, and this model is used to predict the non-sampled values in the population, and hence the finite population total. Inferences are based on the joint distribution of Y and I. Rubin (1976) shows that under probability sampling, inferences can be based on the distribution of Y alone, provided the design variables are included in the model, and the distribution of I given Y is independent of the distribution of Y conditional on the survey design variables.

There are two main variants of the model-based paradigm; frequentist superpopulation modeling and Bayesian modeling. We consider Bayesian modeling (Little, 2004; Little and Zheng, 2007), where we specify a prior distribution for the parameters θ, as well as a distribution for the population values $Y = (Y_S, Y_{S^c})$ conditional on θ, where Y_S denotes the observed values in the sample and Y_{S^c} denotes the unobserved values of the population quantities. Inferences for $Q(Y)$ are based on the posterior predictive distribution of the non-sampled values given the sampled values.

The model-based approach has optimal properties when the model is correctly specified. However, this approach relies on parametric distributional assumptions, and can fail if the model is misspecified. Zheng and Little (2003), Zheng and Little (2005) and Chen et al. (2010) use penalized spline models for estimation and inference and show that such models have good frequentist properties for a variety of populations.

1.2 Problem overview

In populations with differing size units, larger units often contribute more to population quantities of interest than smaller units. A popular design that includes larger units with higher probability is sampling with probability proportional to size (PPS). Specifically, suppose a size measure X is known for all units in the population, and unit t is selected with probability π_t proportional to its size x_t. In particular, we focus on systematic PPS sampling: The procedure first selects a random starting point, and then selects units systematically from a randomly ordered list, at regular intervals on a scale of cumulated sizes. Units that would be selected with probability one are removed and put in a "certainty" stratum (Särndal et al., 2003).

We consider the problem of estimating the population total, $T = \sum_{t=1}^{N} y_t$ of a continuous outcome Y, from a systematic PPS sample of size n. The classical estimator is the Horvitz-Thomson (HT) (Horvitz and Thompson, 1952) estimate

$$\hat{T}_{HT} = \sum_{t=1}^{n} \frac{y_t}{\pi_t} \tag{1}$$

where π_t is the inclusion probability for y_t and the summation is over the n sampled units. Firth and Bennett (1998) show that Eq. (1) can also be obtained in a model-based framework as the projective estimator for a "HT model", namely a linear regression through the origin with residual variance proportional to the square of size. We henceforth refer to such population structure as the "HT population". The HT estimate is design unbiased, but it can be inefficient when the HT model

is not a good approximation to reality. Model-assisted methods such as the generalized regression estimators extend the design-based framework and exploit additional auxiliary variables that are known in the entire population, resulting in improved estimates of the population total (Särndal et al., 2003; Breidt et al., 2005).

Zheng and Little (2003) showed that including the sizes of the non-sampled units as predictors in a spline model can result in improved estimates of the finite population total, without requiring strong parametric assumptions. The penalized spline model has a flexible mean structure, but outcomes in PPS samples are often heteroscedastic, in that the residual variance of Y increases with size. Zheng and Little (2003) assume that the residual variance of the error is proportional to a known function of the sizes. We fit a Bayesian model where the residual variance of Y is proportional to $X^{2\alpha}$, where α is treated as an unknown parameter and assigned a prior distribution. Our Bayesian approach accounts for the additional variance from uncertainty in the parameters without relying on any asymptotic approximations, or requiring additional resampling methods such as the Jackknife (Zheng and Little, 2005).

The HT estimate does not involve the sizes of non-sampled units, and these are usually not included in the data for analysis. Pfeffermann et al. (1998) proposed a quasi randomization-based method of estimating the finite population total in such situations. Their method estimates the total by a HT type estimator where the inclusion probability of each unit is estimated as that unit's response propensity. However, calculation of the propensity scores themselves requires additional covariate information on the sampled units.

Little and Zheng (2007) extended penalized spline estimation to situations where the sizes of the non-sampled units are unavailable. These sizes are predicted by a modified Bayesian Bootstrap (BB) procedure that adjusts for PPS sampling. The missing survey outcomes Y are then predicted using a penalized spline model, fitted via restricted maximum likelihood. When the sizes of non-sampled units are unavailable but their number and average size are known, they applied a ratio adjustment that constrains the BB estimates of X to sum to their known population total. However, this approach is ad-hoc, and it alters the support of the size variable, leading to potentially poor predictions of Y when the relationship between Y and X is nonlinear. We develop here a more principled Bayesian method that replaces the ratio adjustment with posterior screening. Given draws of the sizes of non-sampled units, draws of the non-sampled values of Y (and hence the population total) are obtained from their posterior distribution.

Details of our proposed methods are provided in the next section. In Section 3, we describe a simulation study to compare point and interval estimates of the population total for various simulated populations. We demonstrate our method on a dataset from the U.S Census Bureau in Section 4. Conclusions and directions for future work are presented in Section 5.

2. Methods

Our two-step method is based on factoring the joint distribution of the size variable X and the outcome variable Y into the marginal distribution of X and the conditional distribution of Y given X. In the first step we fit a PPS-adjusted Bayesian bootstrap model on X^S and impute the non-

sampled sizes with posterior draws from the BB model. In the second step, we fit a Bayesian penalized spline model on (X^S, Y^S), and draw the non-sampled survey outcomes from the posterior predictive distribution of Y given X, where we use the imputed values of sizes as predictors. The resulting draw from the posterior distribution of T is the sum of the values y_t for sampled units t plus the predictions, $y_t^{(d)}$ of the non-sampled units; Such regression-based modeling strategies are inspired by ignorability of the sampling design, when conditioning on design variables (Sugden and Smith, 1984; Rubin, 1976). We now describe the Bayesian models for these two steps.

2.1 Constrained Bayesian Bootstrap

The Bayesian bootstrap (BB) (Rubin, 1981; Aitkin, 2008) is the Bayesian analogue of the bootstrap (Efron, 1979). It is operationally and inferentially quite similar to the bootstrap, however philosophically, the bootstrap simulates the sampling distribution of a statistic estimating a parameter, while the BB simulates the posterior distribution of the parameter. The model assumes that only the values of X in the sampled cases have nonzero probability of occurring in the population, but inferences for summaries like means and totals are not sensitive to violations of this assumption.

Specifically, let $\{\tilde{x}_1, \ldots, \tilde{x}_K\}$ be the set of distinct sizes for the sampled units, and let n_k be the number of sampled cases with size $\tilde{x}_k$, $\sum_{k=1}^{K} n_k = n$. We consider these counts to be multinomial with sample size n and probabilities $(\phi_1, \ldots, \phi_K)$, which are assigned a non-informative Haldane prior, i.e $\text{Dirichlet}(0, 0, \ldots, 0)$. The posterior distribution of $\phi = (\phi_1, \ldots, \phi_K)$ is then Dirichlet with parameters $(n_1 - 1, \ldots, n_K - 1)$.

To create draws of the non-sampled values of X, we apply the algorithm of Little and Zheng (2007), which accounts for PPS selection. Let n_k^* be the number of non-sampled cases with size measure $\tilde{x}_k$, with $\sum_{k=1}^{K} n_k^* = N - n$. Under the BB model, the posterior predictive distribution of these counts is multinomial with sample size $N - n$ and probabilities $(\phi_1^*, \ldots, \phi_K^*)$, where by Bayes' rule, $\phi_k^* = c\phi_k(1 - \pi_k)/\pi_k$. The constant c is chosen so that these probabilities sum to 1 and $\pi_k = n\tilde{x}_k/N\bar{X}$ is the selection probability for units with size $\tilde{x}_k$, with $\overline{X}$ being the population mean size. Values of $\phi^{(d)} = (\phi_1^{(d)}, \ldots, \phi_K^{(d)})$ are drawn from their Dirichlet posterior distribution, and the counts of non-sampled cases are drawn as multinomial with probabilities $\phi_k^{(d)*} = c^{(d)}\phi_k^{(d)}(1 - \pi_k)/\pi_k$ obtained by substituting these draws in the above expression.

The population total of X is considered to be known, and is explicitly used in this algorithm. However, due to the stochastic behavior of the posterior draws, the sum of the sampled and drawn sizes often differs from this known total. To fully exploit the information on $T_x = N\bar{X}$, Little and Zheng (2007) applied a ratio adjustment to rescale the predicted sizes to add to the known total. This ad-hoc adjustment leads to predictions of X outside of the support of the data, which we have found in simulations yields biased estimates of T when the relationship between Y and X is not linear. We address this problem by screening the drawn vectors of non-sampled sizes, and selecting only those that yield sums close to the true population total of X. This approach decisively outperformed the ratio adjustment in simulations described in this study.

The screening mechanism draws $a \times B$ vectors from the non-sampled sizes. These B vectors are sorted in terms of their absolute distance from the true value of the population total non-sampled

sizes, $T_{x,ns} = T_x - \sum_{i=1}^{n} x_i$. Denote these sorted vector of non-sampled sizes by $\tilde{x}_{(1)}, \ldots, \tilde{x}_{(B)}$. If the $B+1$st vector yields a sum that is closer in absolute distance than the maximum distance vector, then this vector replaces $\tilde{x}_{(B)}$. This process is repeated until we have compared all the remaining $(a-1) \times B$ draws with the original B draws. The resulting set of draws are the a % draws with closest absolute distance to the true sum of sizes.

2.2 Bayesian Penalized Spline with Heteroscedastic Errors

Splines are often used to fit local polynomials to data. The general setting is that we assume our outcome variable Y is related to our predictor(s) via the relationship

$$y = f(x) + \varepsilon$$

where $f(x)$ is some unspecified smooth function of x. The problem is to estimate f from the (x_i, y_i) pairs $i = 1, \ldots, n$, with the ε_i being independent random variables centered at zero with some known variance structure. Parametric regression models assume stringent functional forms on the structure of $f(x)$. The parametric approach is efficient if the model is correct, but choosing the wrong form for the model can result in bias. Spline functions are more flexible, and less vulnerable to bias from model misspecification.

We fit penalized splines (Ruppert et al., 2003), which belong to the class of regression splines. These models place knots at prespecified locations and model the function f as piecewise-polynomial functions between the knots. In particular for a single predictor x_i, we assume the following mixed effects model:

$$\begin{aligned} y_i|u_i &= \beta_0 + \sum_{j=1}^{p} \beta_j x_i^j + \sum_{k=1}^{q} u_{ik}(x_i - \kappa_k)_+^p + \varepsilon_i, \\ \varepsilon_i &\overset{indep}{\sim} N(0, \sigma_\varepsilon^2 x_i^{2\alpha}), \quad u_i \overset{indep}{\sim} N(0, \sigma_u^2) \end{aligned} \tag{2}$$

where a_+ represents the positive part of a, p represents the degree of the polynomial basis, q determines the number of knots, and κ_k denotes the kth knot. We consider variances of the form $x^{2\alpha}\sigma_\varepsilon^2$ for some constant α, a form that includes a variety of common variance structures for survey data. Smoothness is achieved through the random effects u, which lead to a likelihood that imposes a penalty on the regression coefficients $\beta_0, \ldots, \beta_p$. Letting

$$X = \begin{bmatrix} 1 & x_1 & \ldots & x_1^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \ldots & x_n^p \end{bmatrix}, \quad Z = \begin{bmatrix} (x_1 - \kappa_1)_+^p & \ldots & (x_1 - \kappa_q)_+^p \\ \vdots & \ddots & \vdots \\ (x_n - \kappa_1)_+^p & \ldots & (x_n - \kappa_q)_+^p \end{bmatrix}, \quad D = \begin{bmatrix} \mathbf{0}_{q\times q} & \mathbf{0}_{q\times K} \\ \mathbf{0}_{q\times K} & \mathbf{I}_{K\times K} \end{bmatrix},$$

where $\mathbf{0}$ is the zero matrix and I is the identity matrix, Eq.(2) can be re-expressed in matrix notation as

$$y = X\beta + Z\mathbf{u} + \varepsilon, \quad \text{Cov}\begin{pmatrix} u \\ \varepsilon \end{pmatrix} = \begin{bmatrix} \sigma_u^2 \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma_\varepsilon^2 \Lambda \end{bmatrix}, \tag{3}$$

where $X\beta$ is the pure polynomial component of the spline, Zu is the component with the spline basis functions and $\Lambda = diag({x_i}^{2\alpha}\sigma_\varepsilon^2)$. Using a Lagrange multiplier argument, it can be shown that under the ℓ_2 or ridge penalty, the maximum likelihood estimate of β is $\hat{\beta} = (X'X + \lambda^2 D)^{-1}X'y$ where λ^2 is the tuning parameter controlling the amount of smoothing. Thus, writing $C = [X, Z]$, the fitted values $\tilde{f}$ can then be written as $\tilde{f} = C(C'C + \lambda^2 D)^{-1}C'y$. There is a large literature on methods for choosing the tuning parameter λ^2, however the mixed model approach automatically determines the tuning parameter as the ratio of the two variance components, i.e, $\lambda^2 = \sigma_\varepsilon^2/\sigma_u^2$.

We adopt a Bayesian penalized spline (BPS) approach that simulates the posterior distribution of the parameters, via iterative Markov Chain Monte Carlo (MCMC) methods. The parameters β are assigned a uniform prior and the variance terms are modeled as inverse gamma random variables. Small values for the hyperparameters of σ_ε^2 and σ_u^2, namely A_ε , B_ε , A_u, and B_u are chosen that result in relatively noninformative but finite priors (Ruppert et al., 2003; Chen et al., 2010; Gelman, 2004).

The parameter α allows for heteroscedastic errors and is assigned a uniform prior distribution between -2 and 2, a range which encompasses the set of plausible values in most applications. Draws from the joint posterior of the parameters $(\beta, \mathbf{u}, \sigma_u^2, \sigma_\varepsilon^2, \alpha)$ can then be obtained via a Metropolis-Hastings (MH) within Gibbs sampler. A random-walk MH, with a normal proposal kernel was used for the univariate MH step. Computational details of this algorithm are presented in the Appendix. Algorithm 1 summarizes this method, referred to henceforth as the Bayesian Penalized Spline (BPS) procedure.

Algorithm 1 BPS algorithm

1. (Univariate MH Step) Draw α from:
$$\alpha \sim \left(\prod_{i=1}^n x_i^{-\alpha}\right) e^{-\frac{1}{2\sigma_\varepsilon^2}\left(y - C\begin{pmatrix}\beta\\ \mathbf{u}\end{pmatrix}\right)'\Lambda^{-1}\left(y - C\begin{pmatrix}\beta\\ \mathbf{u}\end{pmatrix}\right)}$$
where $\Lambda = diag(x_i^{2\alpha})$ and $C = [X\ Z]$
2. Sample $(\beta, \mathbf{u})$ from the multivariate normal distribution:
$$(\beta', \mathbf{u}')' \sim \mathbf{N}\left\{\left(C'\Lambda^{-1}C + \frac{\sigma_\varepsilon^2}{\sigma_u^2}D\right)^{-1}C'\Lambda^{-1}y, \sigma_\varepsilon^2\left(C'\Lambda^{-1}C + \frac{\sigma_\varepsilon^2}{\sigma_u^2}D\right)^{-1}\right\}$$
3. Sample σ_u^2 from inverse-gamma distribution:
$$\sigma_u^2 \sim IG\left(A_u + \frac{q}{2}, B_u + \frac{1}{2}\|\mathbf{u}\|^2\right)$$
where A_u, and B_u are the hyperparameters of σ_u^2 and q is the number of knots.
4. Sample σ_ε^2 from inverse-gamma distribution:
$$\sigma_\varepsilon^2 \sim IG\left(A_\varepsilon + \frac{n}{2}, B_\varepsilon + \frac{1}{2}(y - X\beta - Z\mathbf{u})'\Lambda^{-1}(y - X\beta - Z\mathbf{u})\right)$$
where A_ε and B_ε are the hyperparameters of σ_ε^2 and n is the sample size.
5. Return to step 1 and iterate.

2.3 Inference for the finite population total

Combining the Bayesian bootstrap and BPS algorithms presented above, draws from the posterior distribution of T are simulated by a two-step procedure, based on the factorization

$$f(X, Y) = f(X)f(Y|x). \tag{4}$$

The first step of the procedure imputes the sizes of the non-sampled units via the BB algorithm. These imputed sizes are then used as predictors in a BPS model to predict the missing survey outcomes. The finite population total is then estimated as the sum of the actual y's in the sample plus that of the predicted values of y for the non-sampled cases. [1] Multiple draws are generated to simulate the posterior predictive distribution of T_{S^c}, which denotes the sum of the non-sampled y's. Specifically, for each of the B posterior draws of non-sampled sizes, we obtain multiple draws from the posterior predictive distribution of penalized spline predictions. We consider both unconstrained BB predictions, and predictions that are selected to yield a total close to the population total of X. The method resulting from constrained BB (BC) is midway between the Spline method presented in Zheng and Little (2003) and that from the unconstrained BB algorithm (BU) in terms of available information on the size variables. The sum of the non-sampled sizes in the two-step procedure can be estimated via

$$\mathrm{E}(T_{S^c}|data) = \mathrm{E}\left[\mathrm{E}(T_{S^c}|X_{S^c}^{(d)}, data)\right], \tag{5}$$

where $\mathrm{E}(T_{S^c}|data, x_t^{(d)}, data)$, is the Bayes estimate given a vector of posterior draws of non-sampled sizes $X_{S^c}^{(d)}$. The variance of the spline estimator can be directly calculated from the posterior predictive distribution of T, while the variance of the BB-BPS method can be decomposed into a component for estimating the non-sampled sizes and a component from the spline prediction model with known sizes. Using the law of total variation and conditioning on the draws of the size variables obtained from the BB algorithm, we can write

$$\begin{aligned} \mathrm{Var}(T|data) &= \mathrm{Var}(T_{S^c}|data) \\ &= \mathrm{Var}\left[\mathrm{E}(T_{S^c}|X_{S^c}^{(d)}, data)\right] + \mathrm{E}\left[\mathrm{Var}(T_{S^c}|X_{S^c}^{(d)}, data)\right], \end{aligned} \tag{6}$$

where $\mathrm{E}(T|x_t^{(d)}, data)$ and $\mathrm{Var}(T|x_t^{(d)}, data)$ are the posterior mean and variance of the estimated population total from the BPS algorithm, conditional on the sizes drawn from the BB algorithm. The outer terms are the means and variances of the aforementioned over all BB draws. We rely on asymptotic normal theory to construct 95% credible intervals for T.

[1] The same methodology can also be applied to estimating the population mean and yields qualitatively similar results.

3. Simulation Study

3.1 Simulation Design

A simulation study was conducted to assess the above methods. We simulated one hundred distinct values of size x_t from a $\text{Dirichlet}(1, 1, \ldots, 1)$ distribution and replicated each distinct value of x_t 30 times in the simulated populations resulting in populations of size 3000. The support of this Dirichlet distribution is the open 99-dimensional simplex, thus the simulated sizes take extremely small numeric values. We scaled these values by a factor of 100 to avoid numerical instabilities arising from inverting close-to zero values in the HT estimator, as well as preventing singularities in the design matrix corresponding to the pure polynomial component of the spline model. Eighteen populations were considered by drawing the outcome values $Y_t|x_t$ from (i) a normal distribution with mean $f(x_t)$ and constant standard deviation of $\sigma_1 = 0.9$, (ii) a normal distribution with mean $f(x_t)$ and quadratically increasing variances ($\alpha = 1$) of the form ${x_t}^2{\sigma_2}^2$, where $\sigma_2 = 0.8$ and (iii) a log-normal distribution with mean $f(x_t)$ and constant standard deviation of $\sigma_3 = 0.9$ on the logarithm scale. The six different mean functions considered for the survey outcome variables Y are given below.

NULL (No association): $f(x_t) = 1$

LUPO (Linear up through the origin): $f(x_t) = 3.5\, x_t$

LUP (Linear up): $f(x_t) = 5 + 3.5\, x_t$

LDN (Linear down): $f(x_t) = 7 - 3.5\, x_t$

EXP (Exponentially increasing): $f(x_t) = 1 + 0.45\, \exp(0.5 + 1.1\, x_t)$

SINE (Sinusoidal pattern): $f(x_t) = 2 + 5\, \sin(2.5\, x_t)$

Figures 1 and 2 display the normally distributed populations with constant and increasing variances respcetively. Figure 3 depicts the populations with log-normal errors. One thousand samples of size 300 were drawn from each of the above populations by systematic PPS sampling, and the knots were chosen at equally spaced quantiles of the sampled size variables. We conducted simulations with 5, 15, and 30 knots in each PPS sample. The BPS models were fitted using linear as well as cubic spline basis. The following four estimates of the population total T were computed for each sample:

- HT: The Horvitz-Thompson estimator;
- BPS: Sum of sampled values of Y and predictions of non-sampled values from the BPS model. This method assumes the sizes x_t to be known for all units in the population and is used as a benchmark;
- BUBPS: Unconstrained form of the two-step method, described in Section 2.3;

- BCBPS: Constrained form of the two-step method, described in Section 2.3.

In order to assess confidence coverage, we require a method for estimating the variance of HT. We performed a preliminary simulation study comparing several alternative methods for various population structures. Methods compared were the leave-one out jackknife; leave one-out jackknife with a finite population correction for PPS sampling in Wolter (2007); the stratified jackknife method described in Zheng and Little (2005) and Chen et al. (2010) with varying number of strata; the Brewer method implemented in the `survey` package in R (Lumley, 2004), and the with-replacement approximation method in (Zheng and Little, 2005). The stratified jackknife method with 10 strata gave the best results, yielding the narrowest intervals with close to nominal coverage, so we present results for that method here.

For BPS, we estimate T and its variance as the sample mean and variance of the posterior draws of T from BPS. Bayes estimates of T, and the corresponding variance estimates for methods BUBPS, and BCBPS were obtained from Equations (5) and (6) respectively. We performed our analysis using 200 uncorrelated draws from the BPS model, leaving the first 2000 iterations for the burn-in period. Convergence of the chain was assessed via the the Gelman-Rubin statistic (Gelman and Rubin, 1992), which was calculated for each parameter, by running three chains from randomly dispersed starting points.

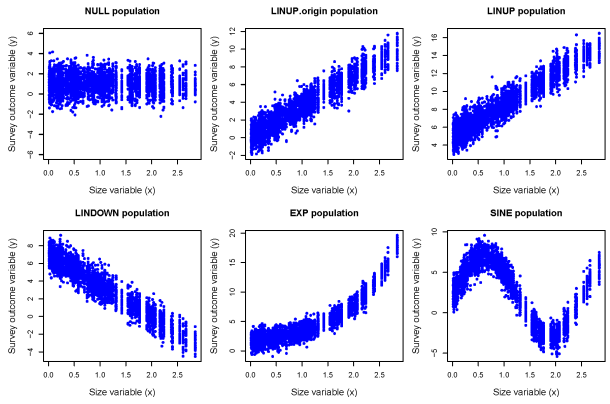

Figure 1: Simulated populations with homoscedastic normal errors (100 unique sizes)

3.2 Simulation Results

Table 1 displays the simulation results, using a spline with 5 equally spaced knots and cubic spline basis. The columns represent the six population structures, NULL, LUPO, LUP, LDN, EXP and SINE, for each of the three different error structures, constant, increasing and lognormal. The rows of the table display the summary measures for the four different estimators.

The last block in the table displays the simulation mean and standard deviation of the posterior mean of α for each population structure (rounded to the second decimal). As we see the posterior

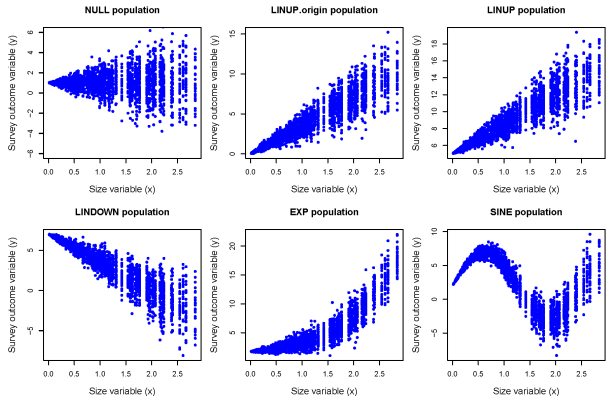

Figure 2: Simulated populations with heteroscedastic normal errors (100 unique sizes)

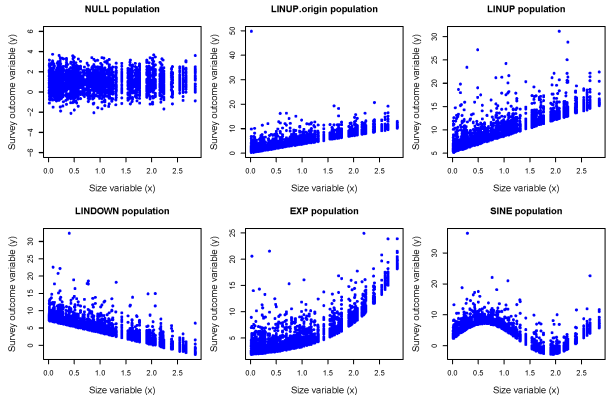

Figure 3: Simulated populations with homoscedastic lognormal errors (100 unique sizes)

mean is centered around the true value of α, even when the normality assumption is violated.

The top three blocks in Table 1 compare the four methods in terms of point estimation: The first block shows the relative root mean square error (as a percentage of the true value of T), the second block gives the relative empirical bias (as a percentage of the true value of T), and the third block displays the empirical relative precision of each estimator with respect to the HT estimator.

As discussed in Zheng and Little (2003), the BPS method, which requires knowledge of all the size variables in the population, yields design-consistent estimates of the population total. BPS yields similar RMSE to HT for data simulated under the HT model, that is the LUPO population with quadratically increasing normal error variances, suggesting that the penalty from fitting a flexible mean function is minor. For other populations, BPS has smaller RMSE than HT, and sometimes the gains are substantial. This replicates the previous findings in Zheng and Little (2003).

Table 1: Summary measures for point estimation (100 unique sizes in the population): Root Mean Square Error (RMSE) as a percentage of true population total, empirical bias as a percentage of true population total, empirical relative precision, $1000\times$ non-coverage of 95% intervals (target value is 50), and average widths of 95% intervals (with respect to the true value of T).

	Population	Constant Error Variance						Increasing Error Variance						Lognormal Error					
	Structure	NULL	LUPO	LUP	LDN	EXP	SINE	NULL	LUPO	LUP	LDN	EXP	SINE	NULL	LUPO	LUP	LDN	EXP	SINE
	BPS	8.79	2.42	0.89	2.35	1.80	3.25	3.74	1.03	0.48	1.28	0.54	1.63	8.71	2.75	2.01	4.03	2.15	3.94
Relative	BCBPS	8.13	2.32	0.88	2.33	1.79	4.49	3.79	1.19	0.52	1.38	0.65	4.07	8.40	2.65	1.95	3.95	2.10	4.16
RMSE	BUBPS	8.14	4.98	2.19	4.99	2.92	3.92	3.75	5.14	2.05	4.76	2.81	2.77	8.28	4.02	2.17	4.55	2.70	3.97
	HT	12.73	2.92	3.55	11.81	3.84	4.76	7.05	1.03	3.23	11.22	2.26	3.64	12.52	8.81	4.48	9.95	4.83	5.94
Relative	BPS	-1.36	-0.39	0.10	-0.08	-0.05	-0.49	0.09	0.04	0.01	-0.02	0.01	-0.12	-0.14	-0.24	0.10	0.06	-0.44	-0.65
empirical	BCBPS	-1.04	-0.04	0.21	-0.40	-0.24	2.78	0.04	0.34	0.13	-0.34	-0.19	3.00	-0.19	0.04	0.20	-0.19	-0.55	1.85
bias	BUBPS	-1.11	1.39	0.80	-1.81	1.12	-0.50	0.04	1.94	0.79	-1.89	1.23	-0.13	-0.17	1.07	0.85	-1.32	0.54	-0.41
	HT	-0.54	-0.02	0.17	0.21	0.14	0.22	0.06	0.05	0.00	-0.03	0.02	-0.03	-0.13	0.21	0.05	-0.09	-0.04	-0.10
Empirical	BPS	2.14	1.50	16.07	25.29	4.52	2.19	3.57	0.98	44.70	76.67	17.77	4.99	2.07	10.33	5.00	6.11	5.26	2.33
relative	BCBPS	2.49	1.57	17.28	26.35	4.64	1.82	3.45	0.82	41.00	70.34	13.44	1.76	2.22	11.03	5.35	6.38	5.66	2.54
precision	BUBPS	2.48	0.37	3.01	6.44	2.02	1.49	3.53	0.05	2.91	6.60	0.80	1.73	2.29	5.17	5.03	5.21	3.32	2.26
	HT	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	BPS	58	74	21	56	46	78	37	10	51	84	1	105	57	73	115	105	63	95
Non-	BCBPS	35	0	0	11	0	0	33	0	0	0	0	0	50	5	35	49	4	2
coverage	BUBPS	12	0	0	0	0	0	22	0	0	0	0	0	24	0	0	0	0	0
	HT	102	76	129	122	159	108	19	8	97	174	59	85	108	122	140	181	158	148
Relative	BPS	33.79	9.07	3.95	9.24	7.57	11.93	16.09	4.43	1.82	4.33	3.57	5.05	32.93	11.62	5.84	12.70	10.32	13.43
average	BCBPS	33.38	14.76	6.21	14.83	12.68	33.27	16.52	12.45	5.15	12.34	10.91	30.77	32.45	14.30	7.17	14.98	12.85	26.09
width	BUBPS	36.91	43.17	17.85	43.09	29.75	37.61	17.05	42.13	17.41	41.79	28.93	34.31	35.74	32.40	15.84	31.69	24.49	29.28
	HT	44.76	10.26	14.81	49.24	12.86	20.50	31.48	5.09	14.12	47.74	10.61	17.54	42.36	14.41	16.72	41.99	16.52	21.97
α	Mean	0.00	0.03	-0.02	0.00	0.01	0.00	1.00	0.99	0.98	1.03	1.01	1.00	0.00	0.00	0.08	0.00	-0.02	0.07
	SD	0.06	0.06	0.06	0.06	0.08	0.08	0.07	0.06	0.08	0.07	0.05	0.07	0.07	0.18	0.22	0.24	0.22	0.21

For most simulated populations BCBPS yielded RMSEs that were only slightly larger than those of BPS, indicating that this method recovers most of the information contained in the sizes of individual non-sampled units. Exceptions are the SINE populations, where BCBPS was markedly inferior to BPS, and in the increasing error variance SINE population, where BCBPS was inferior to both BPS and HT.

The BUBPS method has smaller RMSE than HT in the NULL, LUP, and SINE populations, but considerably higher RMSE in the LUPO populations; results for the LDN population were mixed. In general BUBPS had considerably larger RMSE than BCBPS, reflecting the loss of information from not using the information about total population size. Again the SINE population was an exception to this pattern, with BUBPS having lower RMSE than BCBPS in these cases.

The modeling methods generally yield low empirical bias, similar to HT, although BCBPS had sizeable bias in the SINE populations. Overall, bias from BCBPS was generally smaller than BUBPS, and comparable to that of BPS.

4. Applications

We explore the performance of our method on Washington D.C. housing records obtained from the Public Use Microdata Sample (PUMS) of the U.S. Census Bureau's American Community Survey for the year 2009. It is well known that family income alone does not reflect the spending power of the family members. A crude measure that reflects the spending power is family income per family member. We aim at estimating the population mean of this measure for families in Washington D.C. We overlook the problem of nonresponse in this example and restrict our population to the $1,137$ respondents in the microdata sample. We notice the distribution of number of family members to be right skewed, and therefore chose this variable as a size measure to ensure that the large families are captured in the samples with high probability. Among the respondents, we observed six families with incomes above $\$800,000$ and one eleven-member family. The population size is $N = 1,130$ after the removal of these outliers, and the true value of the population mean is $\overline{Y} = \$46,141$. The number of unique sizes in the population is low, satisfying the underlying assumption of the Bayesian bootstrap procedure. However, the distribution of the survey response variable of interest is highly skewed, violating the inferential assumptions of the penalized spline model. We took 1000 systematic PPS samples from the data with a sampling fraction of 10%. We compare the four methods in terms of point and interval estimates.

Table 2 displays the results for 10 equally spaced knots and linear spline basis. The mean and standard deviation of the Bayes estimates of α are -0.26 and 0.08 respectively across the 1000 iterations. While the HT estimator yields estimates with negligible empirical bias, it fails to achieve nominal coverage. We see that our Bayesian methods yield more efficient estimates. Moreover, we see similar results obtained when assuming that the sizes are known for all units in the population, to when they are imputed using the constrained form of the BB algorithm. Finally, we see that these methods all succeed in achieving nominal coverage.

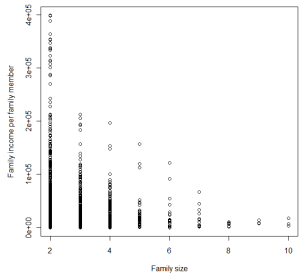

Figure 4: Family income per family member

Table 2: RMSE, empirical bias and average width of CI (as a % of true population mean) and empirical relative precision with respect to the HT estimator.

Summary Measure	BPS	BCBPS	BUBPS	HT
RMSE	7.42	7.41	7.72	9.39
Empirical bias	-3.07	-3.11	-3.31	-0.26
Empirical relative precision	1.94	1.95	1.82	1.00
Avg width of CI	37.04	37.34	39.09	29.97
$1000 \times$ non-coverage	26	20	35	188

5. Discussion

We propose the Bayesian Penalized Spline (BPS) estimator, for inference on the finite population total from a heteroscedastic PPS sample. This framework uses the non-sampled sizes as predictors in the proposed penalized spline model, when the sizes are observed for the entire population, or the imputed sizes, when they are only observed for the sampled units. We consider imputing the non-sampled sizes using constrained and unconstrained Bayesian Bootstrap (BB) models. Unequal probability designs, such as systematic PPS, are popular for their efficiency and ease in administration. However, variance estimation is problematic for the corresponding design-based estimators. The Bayesian predictive approach provides a simulated approximation of the full posterior distribution of the population total, from which variance estimates and credibility intervals are readily computed. This unified approach to inference is particulary desirable for estimators obtained in several steps, as we can easily track all sources of added uncertainty.

The BPS estimator is in general more efficient than the HT estimator. Despite slightly higher empirical bias, the BPS 95% credible intervals provide better coverage and shorter average interval width, especially in the presence of moderate curvature. The HT estimator is known to be optimal for populations where the Y_i/π_i's are exchangeable. However, Bayesian inferences based on the BPS model yield comparative results in such situations, while giving more efficient point estimates for other simulated populations. When the sizes are only recorded for the sampled units, the two-

step estimator that uses imputed non-sampled sizes as predictors in a BPS model still outperforms the HT estimator for populations deviating from the HT structure. Moreover, in our simulated populations other than the SINE, constraining the BB draws to yield a total size similar to the known population value recovers most of the information in the non-sampled sizes. So, conditioning on the total population size is very worthwhile.

The 95% Bayesian credible intervals maintain close to nominal coverage, for populations with limited curvature and normal residual errors. On the other hand, despite their large width, the HT intervals display severe undercoverage when the HT model is incorrect. The two-step methods yield conservative intervals, mainly due to the additional estimated variance from imputation. Furthermore, intervals from the unconstrained BB are more conservative; this is due to the increased imputation variance resulting from extreme configurations in the draws of non-sampled sizes, leading to highly variable estimates of the population total. These extreme configurations also increase the bias, and hence contribute to the reduced efficiency of the unconstrained BB estimates. The constrained BB method eliminates such extreme patterns, by only sampling from the center of the distribution of the sum of the drawn sizes, leading to more efficient estimates and narrower intervals. Hence, even with limited knowledge on the design features, exploiting key external information within a correct model-based framework could lead to superior point and interval estimates.

The BPS model with normal errors produced reasonable estimates of the residual error variance, even in populations with log-normal errors. However, the BPS intervals display slight undercoverage in these settings; Nonetheless, the confidence coverage was better than that of the HT intervals.

For small samples, Bayesian Dirichlet process mixture models with more stringent base distributions may be an appealing alternative to our Bayesian Bootstrap method for imputing non-sampled sizes (Zangeneh et al., 2011). Furthermore, in this paper, we relied on asymptotic normal theory to construct 95% credible intervals for the population total and mean. Constructing intervals for two-step estimators of nonlinear statistics is more delicate, even in large samples, and is the focus of future research.

References

Aitkin, M. (2008). Applications of the bayesian bootstrap in finite population inference. *Journal of Official Statistics*, 24(1):21.

Breidt, F., Claeskens, G., and Opsomer, J. (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika*, 92(4):831–846.

Chen, Q., Elliott, M., and Little, R. (2010). Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling. *Catalogue no. 12-001-X*, page 23.

Cochran, W. (2009). *Sampling techniques*. Wiley-India.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The annals of statistics*, 7(1):1–26.

Firth, D. and Bennett, K. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):3–21.

Gelman, A. (2004). *Bayesian data analysis*. CRC press.

Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Little, R. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99(466):546–556.

Little, R. and Zheng, H. (2007). The Bayesian Approach to the Analysis of Finite Population Surveys. *Bayesian Statistics*, 8(1):1–20.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9(1):1–19.

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

Rubin, D. (1976). Inference and missing data. *Biometrika*, 63(3):581.

Rubin, D. (1981). The bayesian bootstrap. *The Annals of Statistics*, 9(1):130–134.

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric regression*. Cambridge Univ Press.

Särndal, C., Swensson, B., and Wretman, J. (2003). *Model assisted survey sampling*. Springer Verlag.

Sugden, R. and Smith, T. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71(3):495.

Wolter, K. (2007). *Introduction to variance estimation*. Springer Verlag.

Zangeneh, S. Z., Keener, R., and Little, R. J. A. (2011). Bayesian nonparametric estimation of finite population quantities in absence of design information on nonsampled units. Proceedings of the Joint Statistical Meetings.

Zheng, H. and Little, R. (2003). Penalized spline model-based estimation of the finite populations total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19(2):99–118.

Zheng, H. and Little, R. (2005). Inference for the Population Total from Probability-Proportional-to-Size Samples Based on Predictions from a Penalized Spline Nonparametric Model. *Journal of Official Statistics*, 21(1):1–20.