

Bayesian estimation under informative sampling

Terrance D. Savitsky and Daniell Toth

U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. N.E

Washington, D.C. 20212 USA

e-mail: Savitsky.Terrance@bls.gov

Toth.Daniell@bls.gov

Abstract: Bayesian analysis is increasingly popular for use in social science and other application areas where the data are observations from an informative sample. An informative sampling design leads to inclusion probabilities that are correlated with the response variable of interest. Model inference performed on the observed sample taken from the population will be biased for the population generative model under informative sampling since the balance of information in the sample data is different from that for the population. Typical approaches to account for an informative sampling design under Bayesian estimation are often difficult to implement because they require re-parameterization of the hypothesized generating model, or focus on design, rather than model-based, inference. We propose to construct a pseudo-posterior distribution that utilizes sampling weights based on the marginal inclusion probabilities to exponentiate the likelihood contribution of each sampled unit, which weights the information in the sample back to the population. Our approach provides a nearly automated estimation procedure applicable to any model specified by the data analyst for the population and retains the population model parameterization and posterior sampling geometry. We construct conditions on known marginal and pairwise inclusion probabilities that define a class of sampling designs where L_1 consistency of the pseudo posterior is guaranteed. We demonstrate our method on an application concerning the Bureau of Labor Statistics Job Openings and Labor Turnover Survey.

Keywords and phrases: Survey sampling, Gaussian process, Dirichlet process, Bayesian hierarchical models, Latent models, Markov Chain Monte Carlo.

Received July 2015.

1. Introduction

Bayesian formulations are increasingly popular for modeling hypothesized distributions with complicated dependence structures. Their popularity stems from the ease of capturing this dependence by employing models with random effects parameters with a hierarchical construction that regulates the borrowing of information for estimation. Latent parameters are often used in the model to permit flexibility in the estimation of the dependencies among the observations (Dunson, 2010). In social science applications, utilization of latent parameters may be useful for making inference about intrinsic belief states of people from

their observed actions (see for example, Savitsky & Dalal (2013)). Other application areas in which latent parameters may be employed include, engineering and natural science, which use them to parameterize elements of an evolving process.

Data used in these type of applications are often acquired through a complex sample design, resulting in probabilities of inclusion that are associated with the variable of interest. This association could result in an observed data set consisting of units that are not independent and identically distributed. A sampling design that produces a correlation between selection probabilities and observed values is referred to as informative. Failure to account for this dependence caused by the sampling design could bias estimation of parameters that index the joint distribution hypothesized to have generated the population (Holt et al., 1980).

1.1. Examples

We next outline some examples of survey instruments that employ informative sampling designs and associated inferential goals for models estimated on observed samples realized from these surveys.

Example 1: The Survey of Occupational Illnesses and Injuries (SOII) is administered to U.S. business establishments by the U.S. Bureau of Labor Statistics (BLS), in partnership with individual states, in order to capture workplace induced injuries and illnesses. A stratified sampling design is used where strata are indexed by state-industry-size-injury rate. Strata containing establishments that historically express higher injury rates are assigned higher sample inclusion probabilities. The resulting sample will contain a larger proportion of establishments that express higher injury rates than the population, as a whole. States desire to perform regression modeling with variable selection to discover the root causes that predict illnesses and injuries among the population of establishments, estimated from the observed sample. The model-estimated coefficients from the sample will be biased absent correction for over-representation of establishments that tend to express relatively high injury rates.

Example 2: The Current Establishment Statistics (CES) is a BLS survey of U.S. business establishments that collects employment count data across states and industries under a stratified sampling design with strata indexed by the number of employees in each establishment. Strata containing relatively larger establishments are assigned higher inclusion probabilities than those which hold establishments with relatively fewer employees. The distribution of employment totals in the observed sample of establishments will be skewed towards relatively larger values as compared to the population of establishments. An important area of modeling inference is to understand industry-indexed differences in monthly employment trends and correlations among industries in the population. We would use a mixed effects model, parameterized with random effects indexed by industry and month. Estimation of the population distribution under our model from the observed sample will be biased absent some correction for the skewness in the sample towards larger-sized establishments.

Example 3: BLS collects establishment-indexed employment totals in *both* the Quarterly Census of Employment and Wages (QCEW) *and* the CES survey. CES survey participants also provide submissions to the QCEW, such that their reported monthly employment totals for an overlapping time period of interest should be equal between the two instruments, but they are not for approximately 10000 establishments, indicating one or more employment count submission errors for those respondents. A response variable of interest, termed the “error time series”, was created by taking the absolute value of the difference in reported employment totals among the 10000 establishments for each month over a 12 month period. A “response analysis survey” (RAS) of approximately 2000 establishments was taken from this population with the goal to understand the process drivers for committing errors so that BLS may target resources to establishments that mitigate them. The modeling focus is to identify probabilistic clusters of establishments with similar error patterns over the 12 month period and to examine the process by which establishments in each cluster construct their data submissions to BLS. The RAS survey design stratified the population of 10000 establishments based on phenomena of interest expressed in portions of each time series; for example, a big jump in the reported difference at year-end may indicate establishments who count checks that include regular pay and bonuses for each employee, instead of counting employees. Higher inclusion probabilities were assigned to those strata expressing phenomena of relatively greater interest to BLS researchers. Modeling the number of and memberships in probabilistic clusters of error patterns expressed in the population from the RAS sample may be biased because the proportions of error patterns expressed in the sample are designed to be different from the population.

Example 4: The Current Expenditure (CE) survey is administered to U.S. households by BLS for the purpose of determining the amount of spending for a broad collection of goods and service categories and it serves as the main source used to construct the basket of goods later used to formulate the Consumer Price Index. The CE employs a multi-stage sampling design that draws clusters of core-based statistical areas (CBSAs), such as metropolitan and micropolitan areas, from which Census blocks and, ultimately, households are sampled. Economists desire to model the propensity or probability of purchase for a variety of goods and services. The balance of sampled clusters may not be reflective of those in the population; for example, if particularly high income areas are included in the sample. So inference on purchase propensities for the population made from the observed sample will be biased absent correction for the informative sampling design.

Example 5: BLS administers the Job Openings and Labor Turnover survey (JOLTS) to business establishments with the focus to measure labor market dynamics by reporting the number of job openings, hires and separations, which is a leading indicator for employment trends. The sampling design assigns larger inclusion probabilities to establishments with relatively more employees because larger establishments drive the variance in the reported statistics. Our modeling goals are to understand differences in labor force dynamics based on employment ownership (e.g., private, public) and region as part of imputing missing values

with respect to the population generating distribution. As with the CES sampling design, however, our sample will tend to over-represent relatively larger-sized establishments, so that inference and imputation using the sample will be biased for the population. We develop a multivariate count data population generating model in Section 4, where we illustrate the resulting estimation bias from failure to account for the correlations between assigned inclusion probabilities and the response variables of interest for our sample.

The target audience for this article are data analysts who wish to perform some distributional inference using data obtained from an informative sample design on a population using a model they specify, $p(y_i|\boldsymbol{\lambda})$, $\boldsymbol{\lambda} \in \Lambda$, for density, p . We discuss, in the next section, how the limited literature on this topic does not adequately provide a general method for making distributional inference on a population while adjusting for the unequal probabilities of selection.

In this article, we propose an approach that replaces the likelihood with the “pseudo” likelihood (Chambers & Skinner, 2003), $p(y_i|\delta_i = 1, \boldsymbol{\lambda})^{w_i}$, using sampling weight, $w_i \propto 1/\pi_i$. This re-weights the likelihood contribution for each observed unit with intent to re-balance the information in the observed sample to approximate the balance of information in the target finite population; correcting for the informativeness. We show that the proposed method for Bayesian estimation on complex sample data allows for asymptotically consistent inference on any population-generating model specified by the data analyst.

Additionally, this method does not require information about the complex design, other than the probabilities of selection, or about the full population, other than the observed data. We believe this makes the method applicable to more situations. Indeed, it is often the case that the data analyst does not have access to the full design information or auxiliary variables on the population, z_1, \dots, z_N , used to assign the probabilities of selection π_1, \dots, π_N . However, it is common for the probabilities of selection for the units in the sample, π_1, \dots, π_n , to be provided with the observed sample data.

1.2. Review of Methods to Account for Informative Sampling

One current approach is to account for the informativeness by parameterizing the sampling design into the model (Little, 2004). Parameterizing even a simple informative design is often difficult to accomplish and may disrupt desired inference by requiring a change to the underlying population model parameterization. The analyst in Example 3, above, desires to perform inference on an *a priori* unknown clustering of sampled units with their population model for data acquired under a stratified sampling design. Specifying random effects to be indexed by strata will likely conflict with the identification and composition of inferred clusters. Further, the data analyst may not have access to the sampling design, but only indirect information in form of sampling weights. Lastly, the analyst is sometimes required to impute the unobserved units in the finite population, which may be computationally infeasible.

Another approach incorporates the sampling weights into inference about the population, as is our intent, but requires a particular form for the likelihood that

does not allow the analyst to impose their *own* population model formulation of inferential interest. For example, Dong et al. (2014) specifies an empirical likelihood, while Kuniyama et al. (2014) constructs a non-parametric mixture for the likelihood and Rao & Wu (2010) uses a sampling-weighted (pseudo) empirical likelihood. All of these approaches impose Dirichlet distribution priors for the mixture components with hyperparameters specified as a function of the first-order sampling weights. Si et al. (2015) regress the response variable on a Gaussian process function of the weights for sampling designs where subgroups of sampled units have equal weights (e.g., a stratified sampling design). These approaches are designed for inference about simple mean and total statistics, rather than inference for parameters that characterize an analyst-specified population model that is the focus for our proposed method.

One method that uses a plug-in estimator, as do we in our method, is to construct a joint likelihood of the population distribution and sample inclusion in a simple logistic regression model (Malec et al., 1999). This allows one to analytically marginalize over the parameters indexed by the non-sampled units. This approach is limited in application to a class of simple population models that permit analytic integration and may not be applied to more general classes of Bayesian models for the population that we envision in development of our approach.

Perhaps the most general Bayesian approach constructs models to co-estimate parameters for conditional expectations of inclusion probabilities jointly with the population-generating model parameters at each level of a hierarchical construction (Pfeffermann et al., 2006). This formulation is fully Bayesian such that it accounts for all sources of uncertainty in population generation and inclusion of units, but requires a custom implementation of an MCMC sampler for each specified population model, such as their simple two-level linear regression model. The implementations may increase the complexity of the specified model and reduce the quality of posterior mixing in the MCMC, so that they are suitable for relatively simple population probability models.

The method we propose is intended to allow Bayesian inference from *any* population model that may be specified by the the data analyst under an informative sampling design, unlike the alternative methods. It provides asymptotically unbiased estimation using *only* the distribution for the observed sample units and normalized Hájek-like sampling weights. The “plug-in” type method accounts for the informative sampling design by raising the likelihood contribution of each sampled observation to the power of their associated sampling weight. The implementation of the plug-in procedure for Bayesian estimation multiplies the sampling weight into each full conditional log-posterior density. This can then be sampled in the typical sequential scan MCMC.

Unlike these other methods that are prominent in the literature, this method: 1) does not impose a population model (implicitly or explicitly), unlike the most recently-developed methods (Dong et al., 2014; Kuniyama et al., 2014; Rao & Wu, 2010; Si et al., 2015); 2) requires only the sampling weights and does not require parameterizing the sampling design unlike Little (2004); 3) does not require a customized MCMC sampling procedure unlike Pfeffermann

et al. (2006), so can be done automatically; 4) does not require imputing the non-sampled units in the finite population. Our data application and estimation model in the sequel are intended to be representative of common problems for Bayesian inference, and the application data are not readily estimated with these other methods that account for informative sampling.

We formulate the pseudo-posterior density as sampling weight-adjusted plug-in from which we conduct model inference about the population under a dependent, informative sampling design in Section 2. Conditions are constructed that guarantee a frequentist L_1 contraction of the pseudo posterior distribution on the true generating distribution in Section 3. We make an application of the pseudo posterior estimator to construct a regression model for count data using a dataset of monthly job hires and separations collected by the U.S. Bureau of Labor Statistics in Section 4. We reveal large differences for parameter estimates between incorporation versus ignoring the sampling weights. This section also includes a simulation study that compares the pseudo posterior estimated on the observed sample to the posterior estimated on the entire finite population. The paper concludes with a discussion in Section 5. The proofs for the main result, along with two enabling results are contained in an Appendix.

2. Method to account for informative sampling

We begin by constructing the pseudo likelihood and associated pseudo posterior density under any analyst-specified prior formulation on the model, $\boldsymbol{\lambda} \in \Lambda$.

2.1. Pseudo posterior

Suppose there exists a Lebesgue measurable population-generating density, $\pi(y|\boldsymbol{\lambda})$, indexed by parameters, $\boldsymbol{\lambda} \in \Lambda$. Let $\delta_i \in \{0, 1\}$ denote the sample inclusion indicator for units $i = 1, \dots, N$ from the population under sampling without replacement. The density for the observed sample is denoted by, $\pi(y_o|\boldsymbol{\lambda}) = \pi(y|\delta_i = 1, \boldsymbol{\lambda})$, where “ o ” indicates “observed”.

The plug-in estimator for posterior density under the analyst-specified model for $\boldsymbol{\lambda} \in \Lambda$ is

$$\hat{\pi}(\boldsymbol{\lambda}|\mathbf{y}_o, \tilde{\mathbf{w}}) \propto \left[\prod_{i=1}^n p(y_{o,i}|\boldsymbol{\lambda})^{\tilde{w}_i} \right] \pi(\boldsymbol{\lambda}), \quad (1)$$

where $\prod_{i=1}^n p(y_{o,i}|\boldsymbol{\lambda})^{\tilde{w}_i}$ denotes the pseudo likelihood for observed sample responses, \mathbf{y}_o . The joint prior density on model space assigned by the analyst is denoted by $\pi(\boldsymbol{\lambda})$. This pseudo likelihood employs sampling weights, $\{\tilde{w}_i \propto 1/\pi_i\}$, constructed to be inversely proportional to unit inclusion probabilities. Each sampling weight assigns the relative importance of the likelihood contribution for each sample observation to approximate the likelihood for the population. We use $\hat{\pi}$ to denote the noisy approximation to posterior distribution, π , and we make note that the approximation is based on the data, \mathbf{y}_o , and sampling weights, $\{\tilde{\mathbf{w}}\}$, confined to those units *included* in the sample, S .

The total estimated posterior variance is regulated by the sum of the sampling weights. We define unnormalized weights, $\{w_i = 1/\pi_i\}$, and subsequently normalize them, $\tilde{w}_i = \frac{w_i}{\sum_{i=1}^n w_i}$, $i = 1, \dots, n$, to sum to the sample size, n , the asymptotic units of information in the sample. Incorporation of the sampling weights to formulate the pseudo posterior estimator is expected to increase the estimated parameter posterior variances relative to the (unweighted) posterior estimated on a simple random (non-informative) sample because the weights encode the uncertainty with which samples represent the finite population under repeated sampling. This increase in estimated posterior variance may be partly or wholly offset to the extent that the informative sampling design is more efficient than simple random sampling; for example, a stratified sampling design that takes simple random samples within each stratum may produce samples that provide better coverage of the population. Although our method utilizes the weights as a “plug-in”, rather than imposing a prior, Pfeiffermann & Sverchkov (2009) use Bayes rule to demonstrate one may replace the weights with their conditional expectation given the observed response to correct for informative sampling. Replacing the raw weights with their conditional expectation given the observed response may serve to reduce the total variation attributed to weighting (and the resulting posterior uncertainty) in the case where the actual sampled observations express information in different proportions than intended in the sampling design. Even though the conditional *distribution* of the weights given the response is generally different for the observed sample than for the population, nevertheless their conditional *expectations* are equal.

3. Pseudo posterior consistency

We formulate a pseudo posterior distribution in this section and specify conditions under which it contracts on the true generating distribution in L_1 . Let $\nu \in \mathbb{Z}^+$ index a sequence of finite populations, $\{U_\nu\}_{\nu=1, \dots, N_\nu}$, each of size, $|U_\nu| = N_\nu$, such that $N_\nu < N_{\nu'}$, for $\nu < \nu'$, so that the finite population size grows as ν increases. Suppose that $\mathbf{X}_{\nu,1}, \dots, \mathbf{X}_{\nu,N_\nu}$ are independently distributed according to some unknown distribution P , (with density, p) defined on the sample space, $(\mathcal{X}, \mathcal{A})$. If Π is a prior distribution on the model space, $(\mathcal{P}, \mathcal{C})$ to which P is known to belong, then the posterior distribution is given by

$$\Pi(B|\mathbf{X}_1, \dots, \mathbf{X}_{N_\nu}) = \frac{\int_{P \in B} \prod_{i=1}^{N_\nu} \frac{p}{p_0}(\mathbf{X}_i) d\Pi(P)}{\int_{P \in \mathcal{P}} \prod_{i=1}^{N_\nu} \frac{p}{p_0}(\mathbf{X}_i) d\Pi(P)}, \quad (2)$$

for any $B \in \mathcal{C}$, where we refer to $\{\mathbf{X}_{\nu,i}\}_{i=1, \dots, N_\nu}$ as $\{\mathbf{X}_i\}_{i=1, \dots, N_\nu}$ for readability when the context is clear.

Ghosal & van der Vaart (2007) study the rate at which this posterior distribution converges to the assumed true (and fixed) generating distribution P_0 . They prove, under certain conditions on the model space, \mathcal{P} , and the prior distribution, Π , that in P_0 -probability, the posterior distribution concentrates on an arbitrarily small neighborhood of P_0 as $N_\nu \uparrow \infty$.

The observed data on which we focus is not the entire finite population, $\mathbf{X}_1, \dots, \mathbf{X}_{N_\nu}$, but rather a sample, $\mathbf{X}_1, \dots, \mathbf{X}_{n_\nu}$, with $n_\nu \leq N_\nu$, drawn under a sampling design distribution applied to the finite population under which each unit, $i \in (1, \dots, N_\nu)$, is assigned a probability of inclusion in the sample. These unit inclusion probabilities are constructed to depend on the realized finite population values, $\mathbf{X}_1, \dots, \mathbf{X}_{N_\nu}$, at each ν .

3.1. Pseudo posterior distribution

A sampling design is defined by placing a *known* distribution on a vector of inclusion indicators, $\delta_\nu = (\delta_{\nu 1}, \dots, \delta_{\nu N_\nu})$, linked to the units comprising the population, U_ν . The sampling distribution is subsequently used to take an *observed* random sample of size $n_\nu \leq N_\nu$. Our conditions needed for the main result employ known marginal unit inclusion probabilities, $\pi_{\nu i} = \Pr\{\delta_{\nu i} = 1\}$ for all $i \in U_\nu$ and the second-order pairwise probabilities, $\pi_{\nu ij} = \Pr\{\delta_{\nu i} = 1 \cap \delta_{\nu j} = 1\}$ for $i, j \in U_\nu$, which are obtained from the joint distribution over $(\delta_{\nu 1}, \dots, \delta_{\nu N_\nu})$. The dependence among unit inclusions in the sample contrasts with the usual *iid* draws from P . We denote the sampling distribution by P_ν .

Under informative sampling, the marginal inclusion probabilities, $\pi_{\nu i} = P\{\delta_{\nu i} = 1\}$, $i \in (1, \dots, N_\nu)$, are formulated to depend on the finite population data values, $\mathbf{X}_{N_\nu} = (\mathbf{X}_1, \dots, \mathbf{X}_{N_\nu})$. Since the resulting balance of information would be different in the sample, the posterior distribution for $(\mathbf{X}_1 \delta_{\nu 1}, \dots, \mathbf{X}_{N_\nu} \delta_{\nu N_\nu})$, that we employ for inference about P_0 , is not equal to that of Equation 2.

Our task is to perform inference about the population generating distribution, P_0 , using the observed data taken under an informative sampling design. We account for informative sampling by “undoing” the sampling design with the weighted estimator,

$$p^\pi(\mathbf{X}_i \delta_{\nu i}) := p(\mathbf{X}_i)^{\delta_{\nu i} / \pi_{\nu i}}, \quad i \in U_\nu, \quad (3)$$

which weights each density contribution, $p(\mathbf{X}_i)$, by the inverse of its marginal inclusion probability. This construction re-weights the likelihood contributions defined on those units randomly-selected for inclusion in the observed sample ($\{i \in U_\nu : \delta_{\nu i} = 1\}$) to *approximate* the balance of information in U_ν . This approximation for the population likelihood produces the associated pseudo posterior,

$$\Pi^\pi(B | \mathbf{X}_1 \delta_{\nu 1}, \dots, \mathbf{X}_{N_\nu} \delta_{\nu N_\nu}) = \frac{\int_{P \in B} \prod_{i=1}^{N_\nu} \frac{p^\pi}{p_0^\pi}(\mathbf{X}_i \delta_{\nu i}) d\Pi(P)}{\int_{P \in \mathcal{P}} \prod_{i=1}^{N_\nu} \frac{p^\pi}{p_0^\pi}(\mathbf{X}_i \delta_{\nu i}) d\Pi(P)}, \quad (4)$$

that we use to achieve our required conditions for the rate of contraction of the pseudo posterior distribution on P_0 . We recall that both P and δ_ν are random variables defined on the space of measures and possible samples, respectively. Additional conditions are later formulated for the distribution over

samples, P_ν , drawn under the known sampling design, to achieve contraction of the pseudo posterior on P_0 . We assume measurability for the sets on which we compute prior, posterior and pseudo posterior probabilities on the joint product space, $\mathcal{X} \times \mathcal{P}$. For brevity, we use the superscript, π , to denote the dependence on the known sampling probabilities, $\{\pi_{\nu i}\}_{i=1, \dots, N_\nu}$; for example, $\Pi^\pi(B|\mathbf{X}_1\delta_{\nu 1}, \dots, \mathbf{X}_{N_\nu}\delta_{\nu N_\nu}) := \Pi(B|(\mathbf{X}_1\delta_{\nu 1}, \dots, \mathbf{X}_{N_\nu}\delta_{\nu N_\nu}), (\pi_{\nu 1}, \dots, \pi_{\nu N_\nu}))$.

Our main result is achieved in the limit as $\nu \uparrow \infty$, under the countable set of successively larger-sized populations, $\{U_\nu\}_{\nu \in \mathbb{Z}^+}$. We define the associated rate of convergence notation, $\mathcal{O}(b_\nu)$, to denote $\lim_{\nu \uparrow \infty} \frac{\mathcal{O}(b_\nu)}{b_\nu} = 0$.

3.2. Empirical process functionals

We employ the empirical distribution approximation for the joint distribution over population generation and the draw of an informative sample that produces our observed data to formulate our results. Our empirical distribution construction follows Breslow & Wellner (2007) and incorporates inverse inclusion probability weights, $\{1/\pi_{\nu i}\}_{i=1, \dots, N_\nu}$, to account for the informative sampling design,

$$\mathbb{P}_{N_\nu}^\pi = \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} \frac{\delta_{\nu i}}{\pi_{\nu i}} \delta(\mathbf{X}_i), \quad (5)$$

where $\delta(\mathbf{X}_i)$ denotes the Dirac delta function, with probability mass 1 on \mathbf{X}_i and we recall that $N_\nu = |U_\nu|$ denotes the size of the finite population. This construction contrasts with the usual empirical distribution, $\mathbb{P}_{N_\nu} = \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} \delta(\mathbf{X}_i)$, used to approximate $P \in \mathcal{P}$, the distribution hypothesized to generate the finite population, U_ν .

We follow the notational convention of Ghosal et al. (2000) and define the associated expectation functionals with respect to these empirical distributions by $\mathbb{P}_{N_\nu}^\pi f = \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} \frac{\delta_{\nu i}}{\pi_{\nu i}} f(\mathbf{X}_i)$. Similarly, $\mathbb{P}_{N_\nu} f = \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} f(\mathbf{X}_i)$. Lastly, we use the associated centered empirical processes, $\mathbb{G}_{N_\nu}^\pi = \sqrt{N_\nu}(\mathbb{P}_{N_\nu}^\pi - P_0)$ and $\mathbb{G}_{N_\nu} = \sqrt{N_\nu}(\mathbb{P}_{N_\nu} - P_0)$.

The sampling-weighted, (average) pseudo Hellinger distance between distributions, $P_1, P_2 \in \mathcal{P}$, $d_{N_\nu}^{\pi, 2}(p_1, p_2) = \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} \frac{\delta_{\nu i}}{\pi_{\nu i}} d^2(p_1(\mathbf{X}_i), p_2(\mathbf{X}_i))$, where $d(p_1, p_2) = \left[\int (\sqrt{p_1} - \sqrt{p_2})^2 d\mu \right]^{\frac{1}{2}}$ (for dominating measure, μ). We need this empirical average distance metric because the observed (sample) data drawn from the finite population under P_ν are no longer independent. The implication is that our result apply to finite populations generated as *inid* from which informative samples are taken. The associated non-sampling Hellinger distance is specified with, $d_{N_\nu}^2(p_1, p_2) = \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} d^2(p_1(\mathbf{X}_i), p_2(\mathbf{X}_i))$.

3.3. Main result

We proceed to construct associated conditions and a theorem that contain our main result on the consistency of the pseudo posterior distribution under a

class of informative sampling designs at the true generating distribution, P_0 . Our approach extends the main in-probability convergence result of Ghosal & van der Vaart (2007) by adding new conditions that restrict the distribution of the informative sampling design. Suppose we have a sequence, $\xi_{N_\nu} \downarrow 0$ and $N_\nu \xi_{N_\nu}^2 \uparrow \infty$ and $n_\nu \xi_{N_\nu}^2 \uparrow \infty$ as $\nu \in \mathbb{Z}^+ \uparrow \infty$ and any constant, $C > 0$,

(A1) (Local entropy condition – Size of model)

$$\sup_{\xi > \xi_{N_\nu}} \log N(\xi/36, \{P \in \mathcal{P}_{N_\nu} : d_{N_\nu}(P, P_0) < \xi\}, d_{N_\nu}) \leq N_\nu \xi_{N_\nu}^2,$$

(A2) (Size of space)

$$\Pi(\mathcal{P} \setminus \mathcal{P}_{N_\nu}) \leq \exp(-N_\nu \xi_{N_\nu}^2 (2(1 + 2C)))$$

(A3) (Prior mass covering the truth)

$$\Pi\left(P : -P_0 \log \frac{p}{p_0} \leq \xi_{N_\nu}^2 \cap P_0 \left[\log \frac{p}{p_0}\right]^2 \leq \xi_{N_\nu}^2\right) \geq \exp(-N_\nu \xi_{N_\nu}^2 C)$$

(A4) (Non-zero Inclusion Probabilities)

$$\sup_\nu \left[\frac{1}{\min_{i \in U_\nu} \pi_{\nu i}} \right] \leq \gamma, \text{ with } P_0\text{-probability } 1.$$

(A5) (Asymptotic Independence Condition)

$$\limsup_{\nu \uparrow \infty} \max_{i \neq j \in U_\nu} \left| \frac{\pi_{\nu ij}}{\pi_{\nu i} \pi_{\nu j}} - 1 \right| = \mathcal{O}(N_\nu^{-1}), \text{ with } P_0\text{-probability } 1$$

such that for some constant, $C_3 > 0$,

$$N_\nu \sup_\nu \max_{i \neq j \in U_\nu} \left[\frac{\pi_{\nu ij}}{\pi_{\nu i} \pi_{\nu j}} \right] \leq C_3, \text{ for } N_\nu \text{ sufficiently large.}$$

(A6) (Constant Sampling fraction) For some constant, $f \in (0, 1)$, that we term the “sampling fraction”,

$$\limsup_\nu \left| \frac{n_\nu}{N_\nu} - f \right| = \mathcal{O}(1), \text{ with } P_0\text{-probability } 1.$$

Condition (A1) denotes the logarithm of the covering number, defined as the *minimum* number of balls of radius $\xi/36$ needed to cover

$\{P \in \mathcal{P}_{N_\nu} : d_{N_\nu}(P, P_0) < \xi\}$ under distance metric, d_{N_ν} . This condition restricts the growth in the size of the model space, or as noted by Ghosal et al. (2000), the space, \mathcal{P}_{N_ν} , must be not too big in order that the condition specifies an optimal convergence rate (Wong & Shen, 1995). This condition guarantees the existence of test statistics, $\phi_{n_\nu}(\mathbf{X}_1 \delta_{\nu 1}, \dots, \mathbf{X}_{N_\nu} \delta_{\nu N_\nu}) \in (0, 1)$, needed for enabling Lemma B.1, stated in the Appendix, that bounds the expectation of the

pseudo posterior mass assigned on the set $\{P \in \mathcal{P}_{N_\nu} : d_{n_\nu}(P, P_0) \geq \xi_{N_\nu}\}$. Condition (A3) ensures the prior, Π , assigns mass to convex balls in the vicinity of P_0 . Conditions 3.3 and 3.3, together, define the minimum value of ξ_{N_ν} , where if these conditions are satisfied for some ξ_{N_ν} , then they are also satisfied for any $\xi > \xi_{N_\nu}$. Condition (A2) allows, but restricts, the prior mass placed on the uncountable portion of the model space, such that we may direct our inference to an approximating sieve, \mathcal{P}_{N_ν} . This sequence of spaces “trims” away a portion of the space that is not entropy bounded (in condition (A1)). In practice, trimming the space may usually be performed to ensure the entropy bound.

The next three new conditions impose restrictions on the sampling design and associated known distribution, P_ν , used to draw the observed sample data that, together, define a class of allowable sampling designs on which the contraction result for the pseudo posterior is guaranteed. Condition (A4) requires the sampling design to assign a positive probability for inclusion of every unit in the population because the restriction bounds the sampling inclusion probabilities away from 0. Since the maximum inclusion probability is 1, the bound, $\gamma \geq 1$. No portion of the population may be systematically excluded, which would prevent a sample of any size from containing information about the population from which the sample is taken. Condition (A5) restricts the result to sampling designs where the dependence among lowest-level sampled units attenuates to 0 as $\nu \uparrow \infty$; for example, a two-stage sampling design of clusters within strata would meet this condition if the number of population units nested within each cluster (from which the sample is drawn) increases in the limit of ν . Such would be the case in a survey of households within each cluster if the cluster domains are geographically defined and would grow in area as ν increases. In this case of increasing cluster area, the dependence among the inclusion of any two households in a given cluster would decline as the number of households increases with the size of the area defined for that cluster. Condition (A6) ensures that the observed sample size, n_ν , limits to ∞ along with the size of the partially-observed finite population, N_ν .

Theorem 3.1. *Suppose conditions (A1)–(A6) hold. Then for sets $\mathcal{P}_{N_\nu} \subset \mathcal{P}$, constants, $K > 0$, and M sufficiently large,*

$$\mathbb{E}_{P_0, P_\nu} \Pi^\pi \left(P : d_{N_\nu}^\pi(P, P_0) \geq M \xi_{N_\nu} | \mathbf{X}_1 \delta_{\nu 1}, \dots, \mathbf{X}_{N_\nu} \delta_{\nu N_\nu} \right) \leq \frac{16\gamma^2 [\gamma + C_3]}{(Kf + 1 - 2\gamma)^2 N_\nu \xi_{N_\nu}^2} + 5\gamma \exp \left(-\frac{Kn_\nu \xi_{N_\nu}^2}{2\gamma} \right), \quad (6)$$

which tends to 0 as $(n_\nu, N_\nu) \uparrow \infty$.

We note that the rate of convergence is injured for a sampling distribution, P_ν , that assigns relatively low inclusion probabilities to some units in the finite population such that γ will be relatively larger. Samples drawn under a design that expresses a large variability in the sampling weights will express more dispersion in their information similarity to the underlying finite population. Similarly, the larger the dependence among the finite population unit

inclusions induced by P_ν , the higher will be C_3 and the slower will be the rate of contraction.

The separability of the conditions on \mathcal{P} and $\Pi(P)$, on the one hand, from those on the sampling design distribution, P_ν , on the other hand, coupled with the sequential process of taking the observed sample from the finite population reveal that the pseudo posterior, defined on the partially-observed sample from a population, contracts on P_0 through converging to the posterior distribution defined on each fully-observed population. We demonstrate this property of the pseudo posterior in a simulation study conducted in Section 4.1. By contrast, if the posterior distribution, defined on each fully-observed finite population, fails to meet conditions 3.3, 3.3 and 3.3 for the main result from Equation 6, such that it fails to contract on P_0 , then the associated pseudo posterior cannot contract on P_0 , even if the sampling design satisfies conditions (A4), (A5) and (A6).

The proof generally follows that of Ghosal et al. (2000) with substantial modification to account for informative sampling. The L_1 rate of contraction of the pseudo posterior distribution with respect to the joint distribution for population generation and the taking of informative samples is derived. Our approach includes two unique enabling results. Please see Appendix sections A and B for details.

4. Application

We construct a model for count data and perform inference on survey responses collected by the Job Openings and Labor Turnover Survey (JOLTS), introduced in Example 5 of Section 1.1, which is administered by BLS on a monthly basis to a randomly-selected sample from a frame composed of non-agricultural U.S. private (business) and public establishments. JOLTS focuses on the demand side of U.S. labor force dynamics and measures job hires, separations (e.g. quits, layoffs and discharges) and openings. The JOLTS sampling design assigns inclusion probabilities (under sampling without replacement) to establishments to be proportional to the number of employees for each establishment (as obtained from the Quarterly Census of Employment and Wages (QCEW)). This design is informative in that the number of employees for an establishment will generally be correlated with the number of hires, separations and openings. We perform our modeling analysis on a May, 2012 data set of $n = 8595$ responding establishments.

We begin by specifying a finite population regression probability model from which we formulate the sampling-weighted pseudo posterior joint distribution that we use to make inference on model parameters from the population generating distribution with only the observed sample of a finite population. We demonstrate that failing to incorporate sampling weights (e.g. by estimating the posterior distribution defined for the finite population on the observed sample) produces large differences in estimates of parameters.

Our regression model defines a multivariate response as the number of job hires (Hires) for the first response variable and total separations (Seps) as the

second response variable. We construct a single multivariate model (as contrasted with the specification of two univariate models) because these variables of interest tend to be highly correlated such that we expect the regression parameters to express dependence; for example, these two variables are correlated at 60% in our May 2012 dataset.

We formulate a model for count data that accommodates the high degree of over-dispersion expressed in our establishment-indexed multivariate responses due to the large employment size differences across the establishments. Were we working with domain-indexed (e.g., by state or county) responses, we may consider to use a Gaussian approximation for the count data likelihood, but such is not appropriate for us due to the presence of many small-sized establishments. The modeling of count data outcomes is very typical for the analysis of BLS survey data for establishments focused on (un)employment.

We specify the following count data model for the population,

$$y_{id} \stackrel{\text{ind}}{\sim} \text{Pois}(\exp(\psi_{id})) \quad (7)$$

$$\Psi^{N \times D} \sim \mathbf{X}^{N \times D} \mathbf{B}^{P \times D} + \mathcal{N}_{N \times D} \left(\mathbb{I}_N, \Lambda^{-1} \right) \quad (8)$$

$$\mathbf{B} \sim \mathbf{0} + \mathcal{N}_{P \times D} \left(\mathbf{M}^{-1}, [\tau_B \Lambda]^{-1} \right) \quad (9)$$

$$\Lambda \sim \mathcal{W}_D((D+1), \mathbf{I}_D) \quad (10)$$

$$\tau_B \sim \mathcal{G}(1, 1) \quad (11)$$

$$\mathbf{M} \sim \mathcal{W}_P((P+1), \mathbf{I}_P), \quad (12)$$

where $i = 1, \dots, N$ indexes the number of establishments and $d = 1, \dots, D$ indexes the number of dimensions for the multivariate response, \mathbf{Y} . The $N \times D$ log-mean, $\Psi = \begin{pmatrix} \psi'_1, \dots, \psi'_N \end{pmatrix}^{D \times 1}$, may be viewed as a latent response whose columns index the number of job hires (Hires) and total separations (Seps) under our JOLTS application, so that $D = 2$. The number of predictors in the design matrix, \mathbf{X} , is denoted by P and \mathbf{B} are the unknown matrix of population coefficients that serve as the focus for our inference. Our model is formulated as a multivariate Poisson-lognormal model, under which the Gaussian prior of Equation 8 for the logarithm of the Poisson mean allows for over-dispersion (of different degrees) in each of the D dimensions. The priors in Equation 8 and Equation 9 are formulated in matrix variate (or, more generally, tensor product) Gaussian distributions using the notation of Dawid (1981); for example, the prior for the $P \times D$ matrix of coefficients, \mathbf{B} , assigns the $P \times D$ mean $\mathbf{0}$ for a Gaussian distribution that employs a separable covariance structure where the $P \times P$, \mathbf{M} , denotes the precision matrix for the columns of \mathbf{B} , and the $D \times D$, $\tau_B \Lambda$, denotes the precision matrix for the rows. This prior formulation is the equivalent of assigning a PD dimensional Gaussian prior to a vectorization of \mathbf{B} accomplished by stacking its columns with $PD \times PD$ precision matrix,

$\mathbf{M} \otimes (\tau_B \Lambda)$. (See Hoff (2011) for more background). Precision matrices, (\mathbf{M}, Λ) , each receive Wishart priors with hyperparameter values that impose uniform marginal prior distributions on the correlations (Barnard et al., 2000).

We regress the multivariate latent response, Ψ , on predictors representing the logarithm of the overall establishment-indexed number of employees (Emp), obtained from the QCEW, the logarithm of the number of job openings (Open), region (Northeast, South, West, Midwest (Midw)) and ownership type (Private, Federal Government, State Government (State), Local Government (Local)). We convert region and ownership type to binary indicators and leave out the Northeast region and Federal Government ownership to provide the baseline of a full-column rank predictor matrix. We summarize our regression model on the logarithm scale by: $(\psi_{\text{Hires}}, \psi_{\text{Seps}}) \sim 1 + \text{West} + \text{Midw} + \text{South} + \text{State} + \text{Local} + \text{Private} + \log(\text{Emp}) + \log(\text{Opens}) + \text{error}$, where 1 denotes an intercept (Int).

Our population model is hypothesized to generate the finite population of the U.S. non-agricultural establishments, from which we have taken a sample of size $n = 8595$ for May, 2012 as our observations. For ease of reading, we will continue to use \mathbf{Y} and \mathbf{X} , to next define the associated pseudo posterior, though each possesses $n < N$ rows representing the sampled observations, in this context.

The population model likelihood contribution for establishment, i , on dimension, d , is formed with the integration,

$$p(y_{id}|\mathbf{x}_i, \mathbf{B}, \Lambda) = \int_{\mathbb{R}} p(y_{id}|\psi_{id}) \times p(\psi_{id}|\mathbf{x}_i, \mathbf{B}, \Lambda) d\psi_{id}, \quad (13)$$

where sampling weight, $w_i = 1/\pi_i$ and $\tilde{w}_i = n \times w_i / \sum_{i=1}^n w_i$, such that the adjusted weights sum to n , the asymptotic amount of information contained in the sample (under a sampling design that obeys condition (A5)). This integrated likelihood induces the following pseudo likelihood,

$$p^\pi(y_{id}|\mathbf{x}_i, \mathbf{B}, \Lambda) = \left[\int_{\mathbb{R}} p(y_{id}|\psi_{id}) \times p(\psi_{id}|\mathbf{x}_i, \mathbf{B}, \Lambda) d\psi_{id} \right]^{\tilde{w}_i}, \quad (14)$$

which is analytically intractable, so we perform the integration, numerically, in our MCMC using the prior for each ψ_{id} exponentiated by the normalized sampling weight, \tilde{w}_i , which we use to construct its pseudo posterior distribution. Using Bayes rule we present the logarithm of the pseudo posteriors for the latent set of $D \times 1$ log-mean parameters, $\{\boldsymbol{\psi}_i\}$, (which are *a posteriori* independent over $i = 1, \dots, n$), with,

$$\log p^\pi(\boldsymbol{\psi}_i|\mathbf{y}_i, \mathbf{x}_i, \mathbf{B}, \Lambda) \propto \quad (15a)$$

$$\log \left\{ \left[\prod_{d=1}^D \exp(\psi_{id}^{y_{id}}) \exp(-\exp(\psi_{id})) \right]^{\tilde{w}_i} \times \left[\mathcal{N}_D(\boldsymbol{\psi}_i | \mathbf{x}_i' \mathbf{B}, \Lambda^{-1}) \right]^{\tilde{w}_i} \right\} \quad (15b)$$

$$\propto \tilde{w}_i \sum_{d=1}^D [y_{id}\psi_{id} - \exp(\psi_{id})] - \frac{1}{2} (\boldsymbol{\psi}_i - \mathbf{x}_i' \mathbf{B})' \tilde{w}_i \Lambda (\boldsymbol{\psi}_i - \mathbf{x}_i' \mathbf{B}), \quad (15c)$$

where we note in the second expression in Equation 15c that the sampling weights influence the *prior* precision for each $\boldsymbol{\psi}_i$, such that a higher-weighted observation will exert relatively more influence on posterior inference because this observation is relatively more representative of the population. We take samples from the pseudo posterior distribution specified Equation 15c in our MCMC using the elliptical slice sampler of Murray et al. (2010), where we draw $\boldsymbol{\psi}_i \stackrel{\text{ind}}{\sim} \mathcal{N}_D(\mathbf{x}_i' \mathbf{B}, (\tilde{w}_i \Lambda)^{-1})$ and formulate a proposal as a convex combination (parameterized on an ellipse) of this draw from the prior and the value selected on the previous iteration of the MCMC. We evaluate each proposal using the weighted likelihood in the first expression of Equation 15c.

We next illustrate the construction of the pseudo posterior distribution for the $P \times D$ matrix of regression coefficients, \mathbf{B} , (which by D-separation is independent of the observations, (y_{id}) , given (ψ_{id})),

$$p^\pi(\mathbf{B}|\mathbf{Y}, \mathbf{X}, \Psi, \Lambda, \mathbf{M}, \tau_B) \propto \left[\prod_{i=1}^n \mathcal{N}_{n \times D}(\boldsymbol{\psi}_i | \mathbf{B}' \mathbf{x}_i, \mathbb{I}_n, \Lambda^{-1})^{\tilde{w}_i} \right] \times \mathcal{N}_{P \times D}(\mathbf{B} | \mathbf{M}^{-1}, (\tau_B \Lambda)^{-1}) \quad (16a)$$

$$\log p^\pi(\mathbf{B}|\mathbf{Y}, \mathbf{X}, \Psi, \Lambda, \mathbf{M}, \tau_B) \propto \sum_{i=1}^n \left[\frac{\tilde{w}_i}{2} \log |\Lambda| - \frac{\tilde{w}_i}{2} (\boldsymbol{\psi}_i - \mathbf{B}' \mathbf{x}_i)' \Lambda (\boldsymbol{\psi}_i - \mathbf{B}' \mathbf{x}_i) \right] + \log \mathcal{N}_{P \times D}(\mathbf{B} | \mathbf{M}^{-1}, (\tau_B \Lambda)^{-1}). \quad (16b)$$

In a Bayesian setting, the sum of the weights ($n = \sum_{i=1}^n \tilde{w}_i$) impacts the estimated posterior variance as we observe in Equation 16b. We see that weights scale the quadratic product of the Gaussian kernel in Equation 16b such that we may accomplish the same result using the matrix variate formation to define the pseudo likelihood, $\mathcal{N}_{n \times D}(\Psi - \mathbf{X} \mathbf{B} | \tilde{\mathbf{W}}, \Lambda^{-1})$, where $\tilde{\mathbf{W}} = \text{diag}(\tilde{w}_1, \dots, \tilde{w}_n)$, the weights for the sampled observations, from which we compute the following conjugate conditional pseudo posterior distribution defined on the n observations,

$$p^\pi(\mathbf{B}|\mathbf{Y}, \mathbf{X}, \Psi, \Lambda, \mathbf{M}, \tau_B) = \mathbf{h}_B^\pi + \mathcal{N}_{P \times D}(\mathbf{B} | (\phi_B^\pi)^{-1}, \Lambda^{-1}), \quad (17)$$

where $\phi_B^\pi = \mathbf{X}' \tilde{\mathbf{W}} \mathbf{X} + \tau_B \mathbf{M}$ and $\mathbf{h}_B^\pi = (\phi_B^\pi)^{-1} \mathbf{X}' \tilde{\mathbf{W}} \Psi$.

Under employment of a simpler continuous response framework, the conditional posterior for \mathbf{B} retains the same form as Equation 17, except the latent response on the logarithm scale, Ψ , would be replaced by the observed data, \mathbf{Y} . Intuitively, we note using a sampling-weighted pseudo prior for the latent response, Ψ , for sampling coefficients, \mathbf{B} , is analogous to using the sampling-weighted likelihood in the case of an observed, continuous response, \mathbf{Y} .

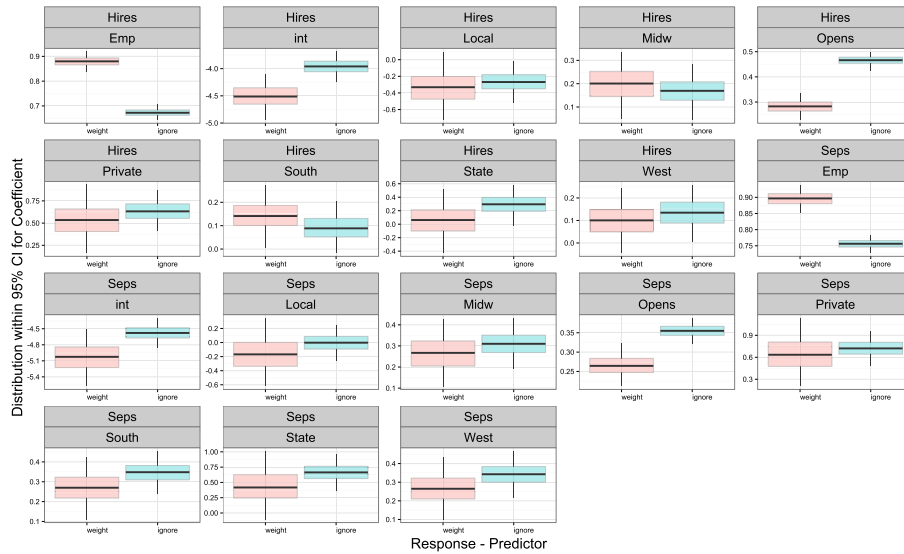


FIG 1. Comparison of posterior densities for the each coefficient in the $(P = 9) \times (D = 2)$ coefficient matrix, \mathbf{B} , within 95% credible intervals, based on inclusion sampling weights in a pseudo posterior (the left-hand plot in each panel) and exclusion of the sampling weights using the posterior distribution defined for the population (in the right-hand plot). Each plot panel is labeled by “predictor, response” for the two included response variables, “Hires”, and “Seps” (total separations).

Each plot panel in Figure 1 compares estimated posterior distributions for a coefficient in \mathbf{B} (within 95% credible intervals), labeled by “predictor, dimension (of the multivariate response)”, when applied to the May, 2012 JOLTS dataset between two estimation models: 1. The left-hand plot in each panel employs the sampling weights to estimate the pseudo posterior for \mathbf{B} , induced by the pseudo posterior for the latent response in Equation 15c; 2. The right-hand plot estimates the coefficients using the posterior distribution defined on the finite population, which may be achieved by replacing $\tilde{\mathbf{W}}$ by the identity matrix to equally weight establishments. Equal weighting of establishments assumes that the sample represents the same balance of information as the population from which it was drawn, which is not the case under an informative sampling design. Comparing estimation results from the pseudo posterior and population posterior distributions provides one method to assess the sensitivity of estimated parameter distributions to the sampling design.

We observe that the estimated results are quite different in both location and variation between estimations performed under the pseudo posterior and population posterior distributions, indicating a high degree of informativeness in the sampling design. The 95% credible intervals for the coefficients of the continuous predictors – (the log of) job openings (Opens) and employment (Emp) – don’t even overlap on both the number of hires (Hires) and separations (Seps)

TABLE 1

Characteristics of single stage, fixed size pps sampling design used in simulation study. n_ν denotes the sample size. CUs denotes the number of certainty units (with inclusion probabilities equal to 1). π_ν denotes the inclusion probabilities (proportional to square root of JOLTS employment), $CV(\pi_\nu)$ denotes the coefficient of variation of π_ν , $Cor(y_{\text{hires}}, \pi_\nu)$ denotes correlation of the number of hires and π_ν and $Cor(y_{\text{seps}}, \pi_\nu)$ denotes the correlation of the number of separations and π_ν .

	n_ν	CUs	$\min(\pi_\nu)$	$\max(\pi_\nu)$	$CV(\pi_\nu)$	$Cor(y_{\text{hires}}, \pi_\nu)$	$Cor(y_{\text{seps}}, \pi_\nu)$
1	500	56	0.02	1.00	2.11	0.80	0.62
2	1000	196	0.04	1.00	1.60	0.69	0.50
3	1500	357	0.07	1.00	1.29	0.61	0.44
4	2500	722	0.14	1.00	0.91	0.51	0.36

responses. The coefficient for the State ownership predictor and the number of hires response is bounded away from 0 when estimated under the (unweighted) population posterior, but is centered on 0 under the sampling-weighted, pseudo posterior. The coefficient posterior variances estimated on the observed sample under the population posterior are understated because they don't reflect the uncertainty with which the information in the sample expresses that in the population (which is captured through the sampling weights).

4.1. Simulation Study

We implement a simulation study to compare the marginal *pseudo* posterior distributions to the (unweighted) population posterior distributions for the regression coefficients, where both are estimated on the observed sample drawn under an informative sampling design. For this study we use the $N = 8595$ observations from the JOLTS May, 2012 data as our *population*. We take 100 Monte Carlo samples of size $\mathbf{n}_\nu = (500, 1000, 1500, 2500)$ establishments using an informative single-stage sample design with unequal inclusion probabilities based on the proportional to size sample used for the real JOLTS survey. Characteristics of the the sampling design, used for this study, at each sample size are presented in Table 1.

This sampling design will induce distributions of the observed samples that will be different from those for the population. The designed correlation between the response and inclusion probabilities will produce observed samples with values skewed towards higher numbers of hires and separations than in the population. Figure 2 demonstrates this difference between the distributions for realized samples under the informative sampling design compared to those for the finite population. The left-most box plot in each of the two panels displays the population distribution for a response value. A single sample is drawn under a sequence of increasing sample sizes for illustration. The next set of box plots displays the resulting distributions for the response values in each sample with size increasing from left-to-right. The left-hand plot panel displays the distributions for the Hires response, while the right-hand panel displays those for the Seps (separations) response variable.

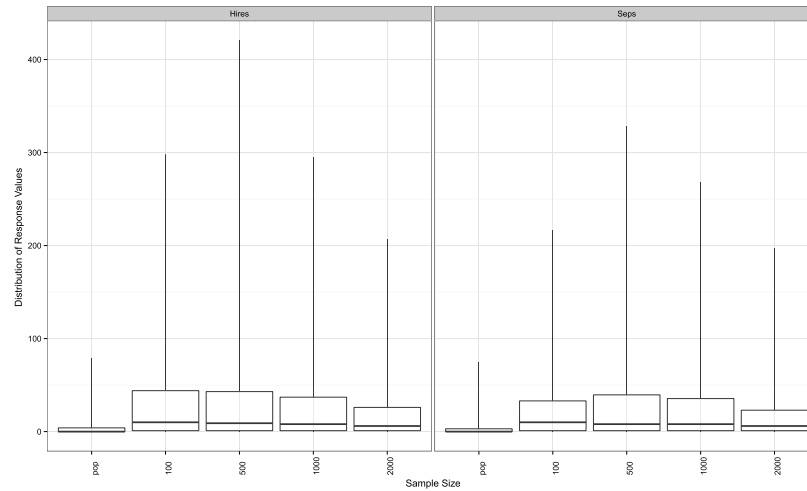


FIG 2. Distributions of response values for population compared to informative samples. The left-most box plot in each of the two plot panels contains the distribution for the JOLTS sample that we use as our “population” in the simulation study. The next set of box plots show the distribution for the response values for increasing sample sizes (from left-to-right) for each sample drawn under our single stage proportion-to-size design. The left-hand plot panel displays the Hires response variable and the right-hand panel displays the Seps (separations) response variable.

Pseudo posterior and population posterior distributions are estimated on each Monte Carlo sample at each sample size in \mathbf{n}_ν . Figure 3 compares estimation of the posterior distribution from the fully-observed population (left-hand box plot) to estimation using the *pseudo* posterior from sample observations taken under the proportional-to-size sampling design. The third box plot in each panel shows the estimation of the posterior distribution estimated on the same sample ignoring the informative sampling design. The last box plot in each panel displays the estimates of the posterior distribution from a simple random sample of the same size, where no correction for the sampling design is required, as a gold standard against which to measure the performance of the pseudo posterior distribution. We estimate the distributions on each of the 100 Monte Carlo draws for each sample size and concatenate the results such that they incorporate both the variation of population generation and repeated sampling from that population. The sample sizes, n_ν , increase from left-to-right across the plot panels. The top set of plot panels display the posterior distributions of the regression coefficient for the employment predictor (Emp) and the hires response (Hires), while the bottom set of panels display the coefficient distributions for the employment predictor (Emp) and the total separations response (Seps).

Scanning from left-to-right in each row across the increasing sample sizes, we readily note a consistent difference in the estimated posterior mean, as expected, between the population model estimated on the samples without ad-

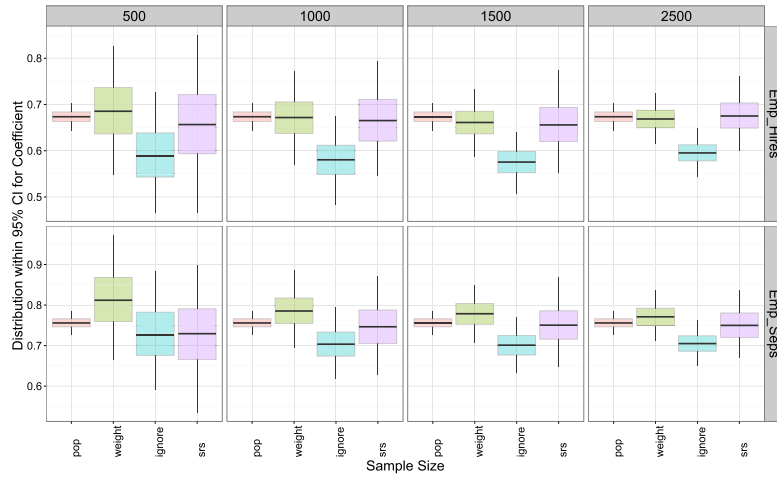


FIG 3. Comparison of posterior densities for 2 coefficients, *Employment-Hires* (top row of plot panels) and *Employment-Separation* (bottom row of plot panels) in **B**, within 95% credible intervals, between estimation on the population (left-hand plot in each panel), estimations from informative samples data taken from that population, which include sampling weights in a pseudo posterior (the second plot from the left in each panel) and exclusion of the sampling weights using the population posterior distribution (the third plot from the left) under a simulation study. The right-most plot presents the posterior density estimated from a simple random sample of the same size for comparison. The simulation study uses the May, 2012 JOLTS sample as the “population” and generates 500 informative samples for a range of sample sizes (of 500, 1000, 1500, 2500, from left-to-right) under a sampling without replacement design with inclusion probabilities set proportionally to the square root of employment levels. A separate estimation is performed on each Monte Carlo sample and the draws from estimated distributions are concatenated over the samples.

justment for the informative sampling design as compared to the mean of the posterior distribution estimated on the entire finite population. The application of the pseudo posterior model, however, produces much less difference (relative to estimation on the fully observed population), though the difference between the estimated pseudo-posterior and the population posterior is yet notably more than that for the simple random sampling result (estimated on samples of the same size as the informative sample). The estimated difference for the pseudo-posterior converges to 0, however, as the sample size increases. The posterior variance for the estimated posterior under simple random sampling remains larger than that for the pseudo posterior estimated on the informative sample because our proportion-to-size sampling design over-samples the highest variance units, which provides better capture of information in the population (which is why this design is used). So, in this case, the improved capture of information in the finite population provided by our sampling design more than overcomes the added variation induced by estimation with the sampling weights. In summary, this simulation study demonstrates the contraction of the pseudo posterior distribution estimated on the sample onto the posterior distribution estimated on a fully-observed finite population.

We were able to directly perform posterior inference about the population using only quantities available for the observed sample under the pseudo posterior full conditional distributions outlined in Section 4. By contrast, Little (2004) offer no modeling approach that parameterizes a proportion-to-size sampling design because they note that each unit is in its own group under the stratum-indexed construction they generally suggest. A typical naive approach, however, is to simply include the sampling weights or a variable highly corrected with them as a predictor *only* for observed units with no imputation of the non-sampled units. This is precisely the construction of the alternative that estimates the population posterior distribution on the informative sample, which is shown as the third box plot in each plot panel of Figure 3, because the employment variable, *Emp*, is included as a predictor and is highly correlated with the sampling weights. This option includes *Emp* only for the sampled units as does the model for the pseudo posterior. The reason for biased inference, even when including a predictor that is highly correlated with sampling weights, is because the distribution for the sampled data conditioned on the sampling weights is not generally equal to the distribution for the population conditioned on the sampling weights by Bayes rule (Pfeffermann & Sverchkov, 2009).

The JOLTS respondent-level data from which samples were drawn for our Monte Carlo simulation study may not be publicly released due to restrictions that protect confidentiality of survey participants. A Monte Carlo simulation study using our pseudo posterior plug-in method may, however, be generated under a Bayesian nonparametric model for functional data that is available from the `growfunctions` package for R (Savitsky, 2015). The package includes a synthetic data generator whose use is illustrated in Savitsky (2014), along with a Monte Carlo simulator that compares parameter estimates when correcting versus ignoring an informative sampling design. Although the functional data model is more complicated than our count data application, the Monte Carlo simulation function available in `growfunctions` produces a figure that demonstrates results very similar to those displayed in Figure 3.

5. Discussion

A variety of broadly applicable approaches are available for incorporating sampling weights into weighted maximum likelihood estimation procedures (Pfeffermann & Sverchkov, 2009) to account for an informative sampling design. Defining easily adaptable algorithms that account for sampling design informativeness under any model for the population specified by the data analyst has proved more challenging for estimating Bayesian probability models. Solutions have focused on parameterizing the sampling design or co-estimating the conditional expectation of inclusion, along with the population-generating model. While these approaches allow estimation using the sampled observations, the implementations typically require a high degree of customization to each population model. We take a different approach that constructs a sample-weighted pseudo-posterior to account for an informative sampling design in our plug-in method that is readily accommodated to any Bayesian population probability model.

We demonstrated the applicability of the plug-in method to a poisson – log-normal model for count data. We showed that the plug-in method reduces estimation bias and posterior estimation includes the uncertainty with which the sample reflects the population on these covariance parameters.

The plug-in method is as easily-implemented and broadly applicable as those methods used for likelihood based optimization. We illustrated in Section 4 that the full conditional posterior distributions defined for the population generating model are easily updated by multiplying the log-likelihoods for $(\{y_{id}\}, \psi_{id})$, by the sampling weights, $\{\tilde{w}_i\}$, without changing the constructions for full conditional posterior distributions. The same concerns that apply in the use of sampling weights under likelihood optimization also apply for Bayesian estimation. The quality of posterior estimation is highly dependent on the population representativeness of the realized sample. Sampling weights may be adjusted based on the composition of the realized sample through estimating the conditional expectation of the weights, given the response values for the observed units. Regressing the weights on the response variables using the observed data and replacing the raw weights with their conditional expectation, known as “weight smoothing”, would be expected to reduce the posterior variances for estimated parameters to the extent that the weights express variance unrelated to the response. While the conditional distribution for the sampling weights given the response under the realized sample is not generally expected to be the same as that for the finite population, their expectations are equal (Pfeffermann & Sverchkov, 2009). We explored such smoothing for our sampling weights for the JOLTS application, but there was little reduction in variance, so we employed the published weights for simplicity.

Even after adjustment, if the composition of the realized sample unevenly reflects information in the population, the weights would express a high variation. Approaches that calibrate the sampling weights to actual population totals, where known, may improve the quality of estimation produced from the plug-in method. BLS performs a calibration adjustment of the JOLTS sampling weights such that the weighted difference of hires and separations reported in the sample ties to the monthly total employment change from the CES survey. (The CES survey is introduced and discussed in Section 1.1). This step adjusts the sampling weights computed from inclusion probabilities under the JOLTS sampling design based on the *actual* sample achieved in each month.

One may, alternatively, implement a more fully Bayesian approach that parameterizes a joint model for the response and sampling weights, specific to a given population generating model, as a method that smoothes the weights in the presence of the response values. Doing so, however, requires imputation of weights and response values for non-sampled units, which can be computationally expensive for a survey that samples from the entire U.S. population of business establishments, as does JOLTS.

Lastly, we construct conditions which, together, define a class of sampling designs under which L_1 consistency of the pseudo posterior is guaranteed. One of these conditions requires that the pairwise sample inclusion dependencies asymptotically decrease to 0. While there are many sampling designs, in prac-

tice, which are members of this class, including the proportion-to-size sampling design used for our JOLTS application, there are some designs which are not – such as a cluster sampling design where the number of clusters grows, but the number of units in each cluster remains relatively fixed. A direction for future research will be to widen the class of allowable designs by incorporating second order (or pairwise) inclusion probabilities for inference, though doing so will introduce some practical restrictions on the specifications for the population generating model.

Appendix A: Proof of Theorem 3.3

Proof. Condition (A1) establishes the existence of test statistics, $\phi_{n_\nu}(X_1\delta_{\nu 1}, \dots, X_{N_\nu}\delta_{\nu N_\nu}) \in (0, 1)$ used to achieve the following result,

$$\begin{aligned} \mathbb{E}_{P_0, P_\nu} \phi_{n_\nu} &\leq \exp(n_\nu \xi_{N_\nu}^2) \cdot \frac{\exp(-Kn_\nu M^2 \xi_{n_\nu}^2)}{1 - \exp(-Kn_\nu M^2 \xi_{N_\nu}^2)} \\ &\leq 2 \exp(-Kn_\nu \xi_{N_\nu}^2), \end{aligned} \quad (18)$$

in Lemmas 2 and 9 of Ghosal & van der Vaart (2007) by setting $\xi = M\xi_{N_\nu}$, and by choosing constant $M > 0$ sufficiently large, such that $KM^2 - 1 > K$.

We will bound the expectation (under (P_0, P_ν) , jointly) of the mass assigned by pseudo posterior distribution for those P at some minimum distance from P_0 ,

$$\begin{aligned} \Pi^\pi(P \in \mathcal{P} : d_{N_\nu}^\pi(P, P_0) \geq M\xi_{N_\nu} | X_1\delta_{\nu 1}, \dots, X_{N_\nu}\delta_{\nu N_\nu}) = \\ \Pi^\pi(P \in \mathcal{P} : d_{N_\nu}^\pi(P, P_0) \geq M\xi_{N_\nu} | X_1\delta_{\nu 1}, \dots, X_{N_\nu}\delta_{\nu N_\nu}) (\phi_{n_\nu} + 1 - \phi_{n_\nu}). \end{aligned} \quad (19)$$

Equation 18 establishes the bound,

$$\mathbb{E}_{P_0, P_\nu} \Pi^\pi(P \in \mathcal{P} : d_{N_\nu}^\pi(P, P_0) \geq M\xi_{N_\nu} | X_1\delta_{\nu 1}, \dots, X_{N_\nu}\delta_{\nu N_\nu}) \phi_{n_\nu} \leq \mathbb{E}_{P_0} \phi_{n_\nu} \leq 2 \exp(-Kn_\nu \xi_{N_\nu}^2), \quad (20)$$

since the pseudo posterior mass is bounded from above by 1. We next enumerate the pseudo posterior distribution for the second term of Equation 19,

$$\begin{aligned} \Pi^\pi(P \in \mathcal{P} : d_{N_\nu}^\pi(P, P_0) \geq M\xi_{N_\nu} | \mathbf{X}_{N_\nu} \boldsymbol{\delta}_{N_\nu}) (1 - \phi_{n_\nu}) = \\ \frac{\int_{P \in \mathcal{P} : d_{N_\nu}^\pi(P, P_0) \geq M\xi_{N_\nu}} \prod_{i=1}^{N_\nu} \frac{p^\pi}{p_0^\pi}(X_i \delta_{\nu i}) d\Pi(P) (1 - \phi_{n_\nu})}{\int_{P \in \mathcal{P}} \prod_{i=1}^{N_\nu} \frac{p^\pi}{p_0^\pi}(X_i \delta_{\nu i}) d\Pi(P)}. \end{aligned} \quad (21)$$

We may bound the denominator of Equation 21 from below, in probability. Define the event,

$$B_{N_\nu} = \left\{ P : -P_0 \log \left(\frac{p}{p_0} \right) \leq \xi_{N_\nu}^2, P_0 \left(\log \frac{p}{p_0} \right)^2 \leq \xi_{N_\nu}^2 \right\}$$

We have from Lemma B.2,

$$\Pr \left\{ \int_{P \in \mathcal{P}} \prod_{i=1}^{N_\nu} \frac{p^\pi}{p_0^\pi} (X_i \delta_{\nu i}) d\Pi(P) \geq \exp [-(1+C)N_\nu \xi^2] \right\} \geq 1 - \frac{\gamma + C_3}{C^2 N_\nu \xi^2},$$

for every $P \in B_{N_\nu}$ and any $C > 0$, $\gamma > 1$, where γ may be set closer to 1 for sampling designs that define a low gradient for inclusion probabilities, $\{\pi_{\nu i}\}$. The constant, $C_3 > 0$, and will be close to 1 for sufficiently large ν . Condition (A3) restricts the prior on B_{N_ν} ,

$$\Pi(B_{N_\nu}) \geq \exp(-N_\nu \xi_{N_\nu}^2 C).$$

Then with probability at least $1 - \frac{16\gamma^2[\gamma+C_3]}{(KM^2f-2\gamma)^2N_\nu\xi^2}$,

$$\begin{aligned} \int_{P \in \mathcal{P}} \prod_{i=1}^{N_\nu} \frac{p^\pi}{p_0^\pi} (X_i \delta_{\nu i}) d\Pi(P) &\geq \exp [-(1+C)N_\nu \xi^2] \Pi(B_{N_\nu}) \\ &\geq \exp(-(1+2C)N_\nu \xi^2) \\ &\geq \exp\left(-\frac{KM^2n_\nu \xi_{N_\nu}^2}{2\gamma}\right), \end{aligned}$$

where we set $1+2C = \frac{KM^2f}{2\gamma}$, where we use condition (A6) to replace $f \times N_\nu$ with n_ν for ν sufficiently large.

Denote this event by,

$$A_{N_\nu}^\pi = \left\{ \int_{P \in \mathcal{P}} \prod_{i=1}^{N_\nu} \frac{p^\pi}{p_0^\pi} (X_i \delta_{\nu i}) d\Pi(P) \geq \exp\left(-\frac{KM^2n_\nu \xi_{N_\nu}^2}{2\gamma}\right) \right\}, \quad (22)$$

such that,

$$\begin{aligned} \Pi^\pi(P \in \mathcal{P} : d_{N_\nu}^\pi(P, P_0) \geq M\xi_{N_\nu} | \mathbf{X}_{N_\nu}, \boldsymbol{\delta}_{N_\nu}) (1 - \phi_{n_\nu}) = \\ \left[\frac{\int_{\{P \in \mathcal{P} : d_{N_\nu}^\pi(P, P_0) \geq M\xi_{N_\nu}\}} \prod_{i=1}^{N_\nu} \frac{p^\pi}{p_0^\pi} (X_i \delta_{\nu i}) d\Pi(P) (1 - \phi_{n_\nu})}{\int_{P \in \mathcal{P}} \prod_{i=1}^{N_\nu} \frac{p^\pi}{p_0^\pi} (X_i \delta_{\nu i}) d\Pi(P)} (\mathbb{I}([A_{N_\nu}^\pi]^c) + \mathbb{I}(A_{N_\nu}^\pi)) \right] \end{aligned}$$

$$\begin{aligned}
&\leq \mathbb{I}([A_{N_\nu}^\pi]^c) + \mathbb{I}(A_{N_\nu}^\pi) \times \left[\exp\left(\frac{KM^2 n_\nu \xi_{N_\nu}^2}{2\gamma}\right) \Pi(\mathcal{P} \setminus \mathcal{P}_{N_\nu}) \right. \\
&\quad \left. + \exp\left(\frac{KM^2 n_\nu \xi_{N_\nu}^2}{2\gamma}\right) \int_{\{P \in \mathcal{P}_{N_\nu} : d_{N_\nu}^\pi(P, P_0) \geq M\xi_{N_\nu}\}} \prod_{i=1}^{N_\nu} \frac{p^\pi}{p_0^\pi} (X_i \delta_{\nu i}) d\Pi(P) (1 - \phi_{n_\nu}) \right]
\end{aligned}$$

Taking the expectation of both sides with respect to the joint distribution, (P_0, P_ν) ,

$$\begin{aligned}
&\mathbb{E}_{P_0, P_\nu} \Pi^\pi(P \in \mathcal{P} : d_{N_\nu}^\pi(P, P_0) \geq M\xi_{N_\nu} | \mathbf{X}_{N_\nu}, \boldsymbol{\delta}_{N_\nu}) (1 - \phi_{n_\nu}) \\
&\leq P([A_{N_\nu}^\pi]^c) + \exp\left(\frac{KM^2 n_\nu \xi_{N_\nu}^2}{2\gamma}\right) \Pi(\mathcal{P} \setminus \mathcal{P}_{N_\nu}) \\
&\quad + \exp\left(\frac{KM^2 n_\nu \xi_{N_\nu}^2}{2\gamma}\right) \times \\
&\quad \mathbb{E}_{P_0, P_\nu} \int_{\{P \in \mathcal{P}_{N_\nu} : d_{N_\nu}^\pi(P, P_0) \geq M\xi_{N_\nu}\}} \prod_{i=1}^{N_\nu} \frac{p^\pi}{p_0^\pi} (X_i \delta_{\nu i}) d\Pi(P) (1 - \phi_{n_\nu}) \\
&\stackrel{(i)}{\leq} \frac{16\gamma^2 [\gamma + C_3]}{(KM^2 f - 2\gamma)^2 n_\nu \xi_{N_\nu}^2} + \exp\left(-\frac{KM^2 n_\nu \xi_{N_\nu}^2}{2\gamma}\right) \\
&\quad + \exp\left(\frac{KM^2 n_\nu \xi_{N_\nu}^2}{2\gamma}\right) \times \\
&\quad \mathbb{E}_{P_0, P_\nu} \int_{\{P \in \mathcal{P}_{N_\nu} : d_{N_\nu}^\pi(P, P_0) \geq M\xi_{N_\nu}\}} \prod_{i=1}^{N_\nu} \frac{p^\pi}{p_0^\pi} (X_i \delta_{\nu i}) d\Pi(P) (1 - \phi_{n_\nu}), \tag{23}
\end{aligned}$$

where in (i) we have used condition 3.3 that bounds from above $\Pi(\mathcal{P} \setminus \mathcal{P}_{N_\nu})$, the prior mass assigned on the portion of the model space that lies outside the sieve, and have plugged in for constant, C .

By conditions (A1), (A4) and Lemma B.1,

$$\begin{aligned}
&\mathbb{E}_{P_0, P_\nu} \int_{\{P \in \mathcal{P}_{N_\nu} : d_{N_\nu}^\pi(P, P_0) \geq M\xi_{N_\nu}\}} \prod_{i=1}^{N_\nu} \frac{p^\pi}{p_0^\pi} (X_i \delta_{\nu i}) d\Pi(P) (1 - \phi_{n_\nu}) \\
&\leq 2\gamma \exp\left(\frac{-KM^2 n_\nu \xi_{N_\nu}^2}{\gamma}\right)
\end{aligned}$$

Returning to the expectation in Equation 23,

$$\begin{aligned}
&\mathbb{E}_{P_0, P_\nu} \Pi^\pi(P \in \mathcal{P} : d_{N_\nu}^\pi(P, P_0) \geq M\xi_{N_\nu} | \mathbf{X}_{N_\nu}, \boldsymbol{\delta}_{N_\nu}) (1 - \phi_{n_\nu}) \\
&\leq \frac{16\gamma^2 [\gamma + C_3]}{(KM^2 - 2\gamma)^2 N_\nu \xi_{N_\nu}^2} + \exp\left(-\frac{KM^2 n_\nu \xi_{N_\nu}^2}{2\gamma}\right) \\
&\quad + \exp\left(\frac{KM^2 n_\nu \xi_{N_\nu}^2}{2\gamma}\right) \times 2\gamma \exp\left(-\frac{KM^2 n_\nu \xi_{N_\nu}^2}{\gamma}\right)
\end{aligned}$$

$$\stackrel{(i)}{\leq} \frac{16\gamma^2 [\gamma + C_3]}{(Kf - 2\gamma)^2 N_\nu \xi_{N_\nu}^2} + 3\gamma \exp\left(-\frac{KM^2 n_\nu \xi_{N_\nu}^2}{2\gamma}\right), \quad (24)$$

where in (i) we use our earlier stated bound, $KM^2 - 1 > K \rightarrow KM^2 > K + 1$.

Bringing all the pieces together,

$$\begin{aligned} & \mathbb{E}_{P_0, P_\nu} \Pi^\pi \left(P \in \mathcal{P} : d_{N_\nu}^\pi(P, P_0) \geq M \xi_{N_\nu} \mid X_1 \delta_{\nu 1}, \dots, X_{N_\nu} \delta_{N_\nu} \right) \\ & \leq 2 \exp(-Kn_\nu \xi_{N_\nu}^2) + \frac{16\gamma^2 [\gamma + C_3]}{(Kf - 2\gamma)^2 N_\nu \xi_{N_\nu}^2} + 3\gamma \exp\left(-\frac{KM^2 n_\nu \xi_{N_\nu}^2}{2\gamma}\right) \\ & \leq \frac{16\gamma^2 [\gamma + C_3]}{(Kf - 2\gamma)^2 N_\nu \xi_{N_\nu}^2} + 5\gamma \exp\left(-\frac{Kn_\nu \xi_{N_\nu}^2}{2\gamma}\right) \end{aligned} \quad (25)$$

where $\gamma \geq 1$ and $C_3 > 0$. The right-hand side of Equation 25 tends to 0 (as $\nu \uparrow \infty$) in P_0 probability. This concludes the proof. \square

Appendix B: Enabling Lemmas

We next construct two enabling results needed to prove Theorem 3.1 to account informative sampling under (A4), (A5) and (A6). The first enabling result, Lemma B.1, extends the applicability of Ghosal & van der Vaart (2007) – Lemmas 2 and 9 for *inid* models to informative sampling without replacement. This result is used to bound from above the numerator for the expectation with respect to the joint distribution for population generation and the taking of the informative sample, (P_0, P_ν) , of the pseudo posterior distribution in Equation 4 on the restricted set of measures, $\{P \in B\}$, where $B = \{P \in \mathcal{P} : d_{N_\nu}(P, P_0) > \delta \xi_{N_\nu}\}$, (for any $\delta > 0$). The restricted set includes those P that are at some *minimum* distance, $\delta \xi_{N_\nu}$, from P_0 under pseudo Hellinger metric, $d_{N_\nu}^\pi$. The second result, Lemma B.2, extends Lemma 8.1 of Ghosal et al. (2000) to bound the probability of the denominator of Equation 4 with respect to (P_0, P_ν) , from below.

Lemma B.1. *Suppose conditions (A1) and (A4) hold. Then for every $\xi > \xi_{N_\nu}$, a constant, $K > 0$, and any constant, $\delta > 0$,*

$$\mathbb{E}_{P_0, P_\nu} \left[\int_{P \in \mathcal{P} \setminus \mathcal{P}_{N_\nu}} \prod_{i=1}^{N_\nu} \frac{p^\pi}{p_0^\pi}(\mathbf{X}_i \delta_{\nu i}) d\Pi(P) (1 - \phi_{n_\nu}) \right] \leq \Pi(\mathcal{P} \setminus \mathcal{P}_{N_\nu}) \quad (26)$$

$$\begin{aligned} & \mathbb{E}_{P_0, P_\nu} \left[\int_{P \in \mathcal{P}_{N_\nu} : d_{N_\nu}^\pi(P, P_0) > \delta \xi} \prod_{i=1}^{N_\nu} \frac{p^\pi}{p_0^\pi}(\mathbf{X}_i \delta_{\nu i}) d\Pi(P) (1 - \phi_{n_\nu}) \right] \leq \\ & 2\gamma \exp\left(\frac{-Kn_\nu \delta^2 \xi^2}{\gamma}\right). \end{aligned} \quad (27)$$

The constant multiplier, $\gamma \geq 1$, defined in condition (A4), restricts the distribution of the sampling design by bounding all marginal inclusion probabilities for population units away from 0. As with the main result, the upper bound is injured by γ .

Proof. We proceed constructively to simplify the form of the expectations on the left-hand side of both Equations 26 and 27 and follow with an application of Lemma 2 (and result 2.2) and Lemma 9 of Ghosal & van der Vaart (2007), which is used to establish the right-hand bound of Equation 27 (based on the existence of tests, ϕ_{n_ν}).

Fixing ν , we index units that comprise the population with, $U_\nu = \{1, \dots, N_\nu\}$. Next, draw a single observed sample of n_ν units from U_ν , indexed by subsequence,

$\{i_\ell \in U_\nu : \delta_{\nu i_\ell} = 1, \ell = 1, \dots, n_\nu\}$. Without loss of generality, we simplify notation to follow by indexing the observed sample, sequentially, with $\ell = 1, \dots, n_\nu$.

We next decompose the expectation under the joint distribution with respect to population generation, P_0 , and the drawing of a sample, P_ν ,

Suppose we draw P from some set $B \subset \mathcal{P}$. By Fubini,

$$\begin{aligned} & \mathbb{E}_{P_0, P_\nu} \left[\int_{P \in B} \prod_{i=1}^{N_\nu} \frac{p^\pi}{p_0^\pi} (\mathbf{X}_i \delta_{\nu i}) d\Pi(P) (1 - \phi_{n_\nu}) \right] \\ & \leq \int_{P \in B} \left[\mathbb{E}_{P_0, P_\nu} \prod_{i=1}^{N_\nu} \frac{p^\pi}{p_0^\pi} (\mathbf{X}_i \delta_{\nu i}) (1 - \phi_{n_\nu}) \right] d\Pi(P) \end{aligned} \quad (28)$$

$$\leq \int_{P \in B} \left\{ \sum_{\delta_\nu \in \Delta_\nu} \mathbb{E}_{P_0} \left[\prod_{\ell=1}^{n_\nu} \left[\frac{p}{p_0} (\mathbf{X}_\ell) \right]^{\frac{1}{\pi_{\nu\ell}}} (1 - \phi_{n_\nu}) \middle| \delta_\nu \right] P_{P_\nu}(\delta_\nu) \right\} d\Pi(P) \quad (29)$$

$$\leq \int_{P \in B} \max_{\delta_\nu \in \Delta_\nu} \mathbb{E}_{P_0} \left[\prod_{\ell=1}^{n_\nu} \left[\frac{p}{p_0} (\mathbf{X}_\ell) \right]^{\frac{1}{\pi_{\nu\ell}}} (1 - \phi_{n_\nu}) \middle| \delta_\nu \right] d\Pi(P) \quad (30)$$

$$\leq \int_{P \in B} \mathbb{E}_{P_0} \left[\prod_{\ell=1}^{n_\nu} \left[\frac{p}{p_0} (\mathbf{X}_\ell) \right]^{\frac{1}{\pi_{\nu\ell}}} (1 - \phi_{n_\nu}) \middle| \delta_\nu^* \right] d\Pi(P) \quad (31)$$

$$\begin{aligned} & \leq \int_{P \in B} \mathbb{E}_{P_0} \left[\prod_{\ell=1}^{n_\nu} \left[\frac{p}{p_0} (\mathbf{X}_\ell) \right] (1 - \phi_{n_\nu}) \middle| \delta_\nu^* \right] d\Pi(P) \\ & \leq \int_{P \in B} P_{\delta_\nu^*} (1 - \phi_{n_\nu}) d\Pi(P), \end{aligned} \quad (32)$$

where $\sum_{\delta_\nu \in \Delta_\nu} P_{P_\nu}(\delta_\nu) = 1$ (Särndal et al., 2003) and

$\delta_\nu^* \in \Delta_\nu = \left\{ \{\delta_{\nu i}^*\}_{i=1, \dots, N_\nu}, \delta_{\nu i}^* \in \{0, 1\} \right\}$ denotes that sample, drawn from the space of all possible samples, Δ_ν , which maximizes the probability under the

population generating distribution for the event of interest. The inequality in Equation 32 results from $\frac{p}{p_0} \leq 1$ and $\frac{1}{\pi_{\nu\ell}} \geq 1$. The conditional expectation of $(1 - \phi_{n_\nu})$ given δ_ν^* is denoted by, $P_{\delta_\nu^*}(1 - \phi_{n_\nu})$.

If $P \in \mathcal{P} \setminus \mathcal{P}_{N_\nu}$,

$$\begin{aligned} \mathbb{E}_{P_0, P_\nu} & \left[\int_{P \in \mathcal{P} \setminus \mathcal{P}_{N_\nu}} \prod_{i=1}^{N_\nu} \frac{p^\pi}{p_0^\pi} (\mathbf{X}_i \delta_{\nu i}) (1 - \phi_{n_\nu}) d\Pi(P) \right] \\ & \leq \int_{P \in \mathcal{P} \setminus \mathcal{P}_{N_\nu}} P_{\delta_\nu^*}(1 - \phi_{n_\nu}) d\Pi(P) \leq \int_{P \in \mathcal{P} \setminus \mathcal{P}_{N_\nu}} d\Pi(P) = \Pi(\mathcal{P} \setminus \mathcal{P}_{N_\nu}), \end{aligned}$$

since $(1 - \phi_{n_\nu}) \leq 1$.

We next establish a bound for $P_{\delta_\nu^*}(1 - \phi_{n_\nu})$ on a sieve or slice. Let $\mathcal{A}_r^\pi = \{P \in \mathcal{P}_{N_\nu} : r\epsilon_{N_\nu} \leq d_{N_\nu}^\pi(P, P_0) \leq 2r\epsilon_{N_\nu}\}$ for integers, r . Under observed $(\mathbf{X}_1 \delta_{\nu 1}^*, \dots, \mathbf{X}_{N_\nu} \delta_{\nu N_\nu}^*) \in \mathcal{X}$, by conditions (A1) and (A4) we have,

$$\sup_{P \in \mathcal{A}_r^\pi} P_{\delta_\nu^*}(1 - \phi_{n_\nu}) \quad (33)$$

$$= \sup_{\{P \in \mathcal{P}_{N_\nu} : r\xi \leq d_{N_\nu}^\pi(P, P_0) \leq 2r\xi\}} P_{\delta_\nu^*}(1 - \phi_{n_\nu}) \quad (34)$$

$$\stackrel{(i)}{\leq} \sup_{\{P \in \mathcal{P}_{N_\nu} : \frac{r\xi}{\sqrt{\gamma}} \leq d_{N_\nu}(P, P_0) \leq \frac{2r\xi}{\sqrt{\gamma}}\}} P_{\delta_\nu^*}(1 - \phi_{n_\nu}) \quad (35)$$

$$\stackrel{(ii)}{\leq} \exp\left(-\frac{Kn_\nu r^2 \xi^2}{\gamma}\right), \quad (36)$$

where the smaller range in (i), $P \in \mathcal{P}_{N_\nu} : \frac{r\xi}{\sqrt{\gamma}} \leq d_{N_\nu}(P, P_0) \leq \frac{2r\xi}{\sqrt{\gamma}}$, increases $P_{\delta_\nu^*}(1 - \phi_{n_\nu})$. The result in (ii) uses condition (A2) to obtain the result of Lemmas 2 and 9 in Ghosal & van der Vaart (2007) where we set $\xi \rightarrow \xi/\sqrt{\gamma}$.

Finally, fixing some value for $\delta > 0$, set $r = 2^\ell \delta$ for a given, for integers, $\ell \geq 0$. Following the approach for bounding the sum over the slices in Wong & Shen (1995), let L be the smallest integer such that $2^{2L} \delta^2 \xi^2 > 2\gamma$, since $d_{N_\nu}^\pi < \sqrt{2\gamma}$ (by our definition of the pseudo Hellinger metric in Section 3.2). Then,

$$\mathbb{E}_{P_{\theta_0}, P_\nu} \left[\int_{\{P \in \mathcal{P}_{N_\nu} : d_{N_\nu}^\pi(P, P_0) \geq \delta\xi\}} \prod_{i=1}^{N_\nu} \frac{p^\pi}{p_0^\pi} (\mathbf{X}_i \delta_{\nu i}) d\Pi(P) (1 - \phi_{n_\nu}) \right] \quad (37)$$

$$= \sum_{\ell=0}^L \mathbb{E}_{P_{\theta_0}, P_\nu} \int_{\{P \in \mathcal{P}_{N_\nu} : 2^\ell \delta \xi \leq d_{N_\nu}^\pi(P, P_0) \leq 2^{\ell+1} \delta \xi\}} \prod_{i=1}^{N_\nu} \frac{p^\pi}{p_0^\pi} (\mathbf{X}_i \delta_{\nu i}) d\Pi(P) (1 - \phi_{n_\nu}) \quad (38)$$

$$\leq \gamma \sum_{\ell=0}^L \exp \left(-\frac{2^{2\ell} K n_\nu \delta^2 \xi^2}{\gamma} \right) \quad (39)$$

$$\leq 2\gamma \exp \left(-\frac{K n_\nu \delta^2 \xi^2}{\gamma} \right), \quad (40)$$

for n_ν sufficiently large such that $\frac{K n_\nu \delta^2 \xi^2}{\gamma} \geq 1$.

This concludes the proof. \square

Lemma B.2. For every $\xi > 0$ and measure Π on the set,

$$B = \left\{ P : -P_0 \log \left(\frac{p}{p_0} \right) \leq \xi^2, P_0 \left(\log \frac{p}{p_0} \right)^2 \leq \xi^2 \right\}$$

under the conditions (A2), (A3), (A4), and (A5), we have for every $C > 0$ and N_ν sufficiently large,

$$Pr \left\{ \int_{P \in \mathcal{P}} \prod_{i=1}^{N_\nu} \frac{p^\pi}{p_0^\pi} (\mathbf{X}_i \delta_{\nu i}) d\Pi(P) \leq \exp [-(1+C)N_\nu \xi^2] \right\} \leq \frac{\gamma + C_3}{C^2 N_\nu \xi^2}, \quad (41)$$

where the above probability is taken with the respect to P_0 and the sampling generating distribution, P_ν , jointly.

The bound of “1” in the numerator of the result for Lemma 8.1 of Ghosal et al. (2000), is replaced with $\gamma + C_3$ for our generalization of this result in Equation 41. The sum of positive constants, $\gamma + C_3$, is greater than 1 and will be larger for sampling designs where the inclusion probabilities, $\{\pi_{\nu i}\}$, express relatively higher gradients. Observing each finite population in a skewed fashion through the taking of an informative sample may only slow the rate of posterior contraction (as compared to contraction of the posterior distribution defined on the *fully* observed finite population).

Proof. By Jensen’s inequality,

$$\begin{aligned} \log \int_{P \in \mathcal{P}} \prod_{i=1}^{N_\nu} \frac{p^\pi}{p_0^\pi} (\mathbf{X}_i \delta_{\nu i}) d\Pi(P) &\geq \sum_{i=1}^{N_\nu} \int_{P \in \mathcal{P}} \log \frac{p^\pi}{p_0^\pi} (\mathbf{X}_i \delta_{\nu i}) d\Pi(P) \\ &= N_\nu \cdot \mathbb{P}_{N_\nu} \int_{P \in \mathcal{P}} \log \frac{p^\pi}{p_0^\pi} d\Pi(P), \end{aligned}$$

where we recall that the last equation denotes the empirical expectation functional taken with respect to the joint distribution over population generating and informative sampling. By Fubini,

$$\mathbb{P}_{N_\nu} \int_{P \in \mathcal{P}} \log \frac{p^\pi}{p_0^\pi} d\Pi(P) = \int_{P \in \mathcal{P}} \left[\mathbb{P}_{N_\nu} \log \frac{p^\pi}{p_0^\pi} \right] d\Pi(P)$$

$$\begin{aligned}
&= \int_{P \in \mathcal{P}} \left[\mathbb{P}_{N_\nu} \frac{\delta_\nu}{\pi_\nu} \log \frac{p}{p_0} \right] d\Pi(P) \\
&= \int_{P \in \mathcal{P}} \left[\mathbb{P}_{N_\nu}^\pi \log \frac{p}{p_0} \right] d\Pi(P) \\
&= \mathbb{P}_{N_\nu}^\pi \int_{P \in \mathcal{P}} \log \frac{p}{p_0} d\Pi(P),
\end{aligned}$$

where we, again, apply Fubini.

Then, the probability statement in the result of Equation 41 is bounded (from above) by,

$$\begin{aligned}
&\Pr \left\{ \mathbb{G}_{N_\nu}^\pi \int_{P \in \mathcal{P}} \log \frac{p}{p_0} d\Pi(P) \leq -\sqrt{N_\nu} \xi^2 (1+C) - \sqrt{N_\nu} P_0 \int_{P \in \mathcal{P}} \log \frac{p}{p_0} d\Pi(P) \right\} \\
&= \Pr \left\{ \mathbb{G}_{N_\nu}^\pi \int_{P \in \mathcal{P}} \log \frac{p}{p_0} d\Pi(P) \leq -\sqrt{N_\nu} \xi^2 (1+C) - \sqrt{N_\nu} \int_{P \in \mathcal{P}} P_0 \log \frac{p}{p_0} d\Pi(P) \right\} \\
&= \Pr \left\{ \mathbb{G}_{N_\nu}^\pi \int_{P \in \mathcal{P}} \log \frac{p}{p_0} d\Pi(P) \leq -\sqrt{N_\nu} \xi^2 (1+C) + \sqrt{N_\nu} \xi^2 = -\sqrt{N_\nu} \xi^2 C \right\},
\end{aligned}$$

where we have again applied Fubini in the second inequality and also the bound for $P_0 \log \frac{p}{p_0} \leq \xi^2$ for P on the set B .

We now apply Chebyshev and Jensen's inequality to bound the probability,

$$\begin{aligned}
&\Pr \left\{ \mathbb{G}_{N_\nu}^\pi \int_{P \in \mathcal{P}} \log \frac{p}{p_0} d\Pi(P) \leq -\sqrt{N_\nu} \xi^2 C \right\} \\
&\leq \frac{\text{Var} \left[\int_{P \in \mathcal{P}} \mathbb{G}_{N_\nu}^\pi \log \frac{p}{p_0} d\Pi(P) \right]}{N_\nu \xi^4 C^2} \tag{42a}
\end{aligned}$$

$$\leq \frac{\int_{P \in \mathcal{P}} \left[\text{Var} \left(\mathbb{G}_{N_\nu}^\pi \log \frac{p}{p_0} \right) \right] d\Pi(P)}{N_\nu \xi^4 C^2} \tag{42b}$$

$$\leq \frac{\int_{P \in \mathcal{P}} \left[\mathbb{E}_{P_0, P_\nu} \left(\mathbb{G}_{N_\nu}^\pi \log \frac{p}{p_0} \right)^2 \right] d\Pi(P)}{N_\nu \xi^4 C^2} \tag{42c}$$

$$\leq \frac{\int_{P \in \mathcal{P}} \left[\mathbb{E}_{P_0, P_\nu} \left(\sqrt{N_\nu} \mathbb{P}_{N_\nu}^\pi \log \frac{p}{p_0} \right)^2 \right] d\Pi(P)}{N_\nu \xi^4 C^2}, \tag{42d}$$

where $\mathbb{E}_{P_0, P_\nu}(\cdot)$ denotes the expectation with respect to the joint distribution over population generation and sampling (from that population) without replacement. We apply Jensen's inequality in Equation 42b and use $\mathbb{E}(X^2) > \text{Var}(X)$ in the third inequality, stated in Equation 42c, and drop the centering term in Equation 42d. We now bound the expectation inside the square brackets on the right-hand side of Equation 42d, which is taken with respect to this joint distribution. In the sequel, define $\mathcal{A}_\nu = \sigma(\mathbf{X}_1, \dots, \mathbf{X}_{N_\nu})$ as the sigma field of information potentially available for the N_ν units in population, U_ν .

$$\begin{aligned}
& \mathbb{E}_{P_0, P_\nu} \left(\sqrt{N_\nu} \mathbb{P}_{N_\nu}^\pi \log \frac{p}{p_0} \right)^2 \\
&= \frac{1}{N_\nu} \sum_{i, j \in U_\nu} \mathbb{E}_{P_0, P_\nu} \left(\frac{\delta_{\nu i} \delta_{\nu j}}{\pi_{\nu i} \pi_{\nu j}} \log \frac{p}{p_0}(\mathbf{X}_i) \log \frac{p}{p_0}(\mathbf{X}_j) \right) \\
&= \frac{1}{N_\nu} \sum_{i \in U_\nu} \mathbb{E}_{P_0} \left[\mathbb{E}_{P_\nu} \left\{ \left(\frac{\delta_{\nu i}}{\pi_{\nu i}^2} \left(\log \frac{p}{p_0}(\mathbf{X}_i) \right)^2 \right) \middle| \mathcal{A}_\nu \right\} \right] \\
&+ \frac{1}{N_\nu^2} \sum_{i \neq j \in U_\nu} \mathbb{E}_{P_0} \left[\frac{\mathbb{E}_{P_\nu} [\delta_{\nu i} \delta_{\nu j} | \mathcal{A}_\nu]}{\pi_{\nu i} \pi_{\nu j}} \log \frac{p}{p_0}(\mathbf{X}_i) \log \frac{p}{p_0}(\mathbf{X}_j) \right] \\
&= \frac{1}{N_\nu} \sum_{i \in U_\nu} \mathbb{E}_{P_0} \left[\left(\frac{1}{\pi_{\nu i}} \right) \left(\log \frac{p}{p_0}(\mathbf{X}_i) \right)^2 \right] \\
&+ \frac{1}{N_\nu} \sum_{i \neq j \in U_\nu} \mathbb{E}_{P_0} \left[\frac{\pi_{\nu ij}}{\pi_{\nu i} \pi_{\nu j}} \log \frac{p}{p_0}(\mathbf{X}_i) \log \frac{p}{p_0}(\mathbf{X}_j) \right] \\
&\leq \xi^2 \sup_\nu \left[\frac{1}{\min_{i \in U_\nu} \pi_{\nu i}} \right] + \xi^2 (N_\nu - 1) \sup_\nu \max_{i \neq j \in U_\nu} \left[\left| \frac{\pi_{\nu ij}}{\pi_{\nu i} \pi_{\nu j}} \right| \right] \\
&\leq \xi^2 (\gamma + C_3),
\end{aligned}$$

for sufficiently large N_ν , where we have applied the condition for $P \in B$ for the first term of the last two inequalities and conditions (A4) and (A5) for the last inequality. We additionally note that $\pi_{\nu ij} = \pi_{\nu j}$ when $i = j$, $i, j \in U_\nu$. This concludes the proof. \square

References

- Barnard, J., McCulloch, R. & Meng, X.-L. (2000), 'Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage', *Statistica Sinica* **10**(4), 1281–1311.
- Breslow, N. E. & Wellner, J. A. (2007), 'Weighted likelihood for semiparametric models and two-phase stratified samples, with application to cox regression',

- Scandinavian Journal of Statistics* **34**(1), 86–102. URL: <http://EconPapers.repec.org/RePEc:bla:scjsta:v:34:y:2007:i:1:p:86-102>
- Chambers, R. & Skinner, C. (2003), *Analysis of Survey Data*, Wiley Series in Survey Methodology, Wiley. URL: <http://books.google.com/books?id=4pYGz69d-LkC>
- Dawid, A. (1981), ‘Some matrix-variate distribution theory: Notational considerations and a Bayesian application’, *Biometrika* **68**(1), 265–274.
- Dong, Q., Elliott, M. R. & Raghunathan, T. E. (2014), ‘A nonparametric method to generate synthetic populations to adjust for complex sampling design features’, *Survey Methodology* **40**(1), 29–46.
- Dunson, D. B. (2010), ‘Nonparametric bayes applications to biostatistics’, *Bayesian nonparametrics* **28**, 223–273.
- Ghosal, S., Ghosh, J. K. & Vaart, A. W. V. D. (2000), ‘Convergence rates of posterior distributions’, *Ann. Statist.* pp. 500–531.
- Ghosal, S. & van der Vaart, A. (2007), ‘Convergence rates of posterior distributions for noniid observations’, *Ann. Statist.* **35**(1), 192–223. URL: <http://dx.doi.org/10.1214/009053606000001172>
- Hoff, P. D. (2011), ‘Separable covariance arrays via the tucker product, with applications to multivariate relational data’, *Bayesian Anal.* **6**(2), 179–196. URL: <http://dx.doi.org/10.1214/11-BA606>
- Holt, D., Smith, T. & Winter, P. (1980), ‘A nonparametric method to generate synthetic populations to adjust for complex sampling design features’, *Journal of the Royal Statistical Society. Series A (General)* **143**, 474–487.
- Kunihamma, T., Herring, A. H., Halpern, C. T. & Dunson, D. B. (2014), Nonparametric bayes modeling with sample survey weights, Technical report, Submitted to Biometrika.
- Little, R. J. (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.
- Malec, D., Davis, W. W. & Cao, X. (1999), ‘Model-based small area estimates of overweight prevalence using sample selection adjustment’, *Statistics in Medicine* **18**, 3189–3200.
- Murray, I., Adams, R. P. & MacKay, D. J. (2010), ‘Elliptical slice sampling’, *JMLR: W&CP* **9**, 541–548.
- Pfeffermann, D., Da Silva Moura, F. A. & Do Nascimento Silva, P. L. (2006), ‘Multi-level modelling under informative sampling’, *Biometrika* **93**(4), 943–959.
- Pfeffermann, D. & Sverchkov, M. (2009), Inference under informative sampling, in D. Pfeffermann & C. Rao, eds, ‘Handbook of statistics 29B: sample surveys: inference and analysis’, Elsevier Science Ltd., pp. 455–487.
- Rao, J. N. K. & Wu, C. (2010), ‘Bayesian pseudo-empirical-likelihood intervals for complex surveys’, *Journal of the Royal Statistical Society Series B* **72**(4), 533–544. URL: <http://EconPapers.repec.org/RePEc:bla:jorssb:v:72:y:2010:i:4:p:533-544>
- Särndal, C.-E., Swensson, B. & Wretman, J. (2003), ‘Model assisted survey sampling (springer series in statistics)’.

- Savitsky, T. (2015), *growfunctions: Bayesian Non-Parametric Dependent Models for Time-Indexed Functional Data*. R package version 0.12. URL: <https://CRAN.R-project.org/package=growfunctions>
- Savitsky, T. D. (2014), ‘Bayesian Non-parametric Mixture Estimation for Time-indexed Functional Data in R’, *To appear in Journal of Statistical Software*.
- Savitsky, T. D. & Dalal, S. R. (2013), ‘Bayesian non-parametric analysis of multirater ordinal data, with application to prioritizing research goals for prevention of suicide’, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. URL: <http://dx.doi.org/10.1111/rssc.12049>
- Si, Y., Pillai, N. S. & Gelman, A. (2015), ‘Bayesian nonparametric weighted sampling inference’, *Bayesian Anal.* **10**(3), 605–625. URL: <http://dx.doi.org/10.1214/14-BA924>
- Wong, W. H. & Shen, X. (1995), ‘Probability inequalities for likelihood ratios and convergence rates of sieve mles’, *Ann. Statist.* **23**(2), 339–362. URL: <http://dx.doi.org/10.1214/aos/1176324524>