Bayesian Uncertainty Estimation Under Complex Sampling

Matthew R. Williams
and
Terrance D. Savitsky
U.S. Bureau of Labor Statistics

August 1, 2018

Abstract

Multistage sampling designs utilized by federal statistical agencies are typically constructed to maximize the efficiency of the target domain level estimator (e.g., indexed by geographic area) within cost constraints to administer survey instruments. Sampling designs are usually constructed to be informative, whereby inclusion probabilities are computed to be correlated with the response variable of interest to minimize the variance of the resulting estimator. Multistage sampling designs may induce dependence between the sampled units; for example, employment of a sampling step that selects geographically-indexed clusters of units in order to efficiently manage the cost of collection. A data analyst may use a sampling-weighted pseudo-posterior distribution to estimate the population model on the observed sample. The dependence induced between co-clustered units inflates the scale of the resulting pseudo-posterior covariance matrix that has been shown to induce under coverage of the credibility sets. While the pseudo-posterior distribution contracts on the true population model parameters (under mild conditions on model complexity and the sampling design), we demonstrate that the scale and shape of the asymptotic distributions are different between each of the MLE, the pseudo-posterior and the MLE under simple random sampling. Motivated by the different forms of the asymptotic covariance matrices and the within cluster dependence, we devise a correction applied as a simple and
fast post-processing step to our MCMC draws taken from the pseudo-posterior distribution. Our updating step projects the pseudo-posterior covariance matrix such that the nominal coverage is approximately achieved with credibility sets that account for both the distributions for population generation, $P_{\theta_0}$, and the multistage, informative sampling, $P_\nu$. We demonstrate the efficacy of our scale and shape projection procedure on synthetic data and make an application to the National Survey on Drug Use and Health.

*Keywords:* pseudo-posterior distribution, Credible set, Cluster sampling, Multistage sampling, Survey sampling, Sampling weights, Markov Chain Monte Carlo.
1 Introduction

Our set-up focuses on the task of the data analyst to estimate a Bayesian model, $P_{\theta_0}$, that they suppose generates values for a random variable, $Y$, for units of a population, $U = (1, \ldots, N)$, from an observed sample, $S = (1, \ldots, n \leq N)$, drawn from that population under a multistage sampling design governed by distribution, $P_{\nu}$. Multistage sampling designs focus to achieve efficient (low variance) estimation of a desired simple quantile, mean or total estimator for a collection of domains within constraints on cost to administer the survey. A first stage of the sampling design often collects contiguous geographic areas from the population into clusters, where a subset of the clusters are randomly selected into the sample for this first stage. The contiguity of areas within each cluster is defined for convenience and cost to collect the sample, but it induces a dependence among units nested in areas within each cluster. Dependencies among sampled units may be additionally promulgated through the drawing of a fixed-sized sample, without replacement, in any stage of the sampling design; for example, by constructing a systematic sampling step with a fixed interval using a random starting point.

The sampling design distribution is induced by specifying marginal inclusion probabilities at each stage. Survey agencies, such as Federal statistical agencies, publish marginal inclusion probabilities for last-stage sampled units, $\pi_i = Pr \{\delta_i = 1\} \in (0, 1]$, for (observed) units, $i = 1, \ldots, n$, sampled in the last stage of the sampling design, where $n$ denotes the number of units in the observed sample and $\delta_i \in \{0, 1\}$ specifies a unit inclusion indicator.

Efficiency of the population estimator, $g(Y)$, is enhanced through designing the inclusion probabilities, $(\pi_i)_{i=1,\ldots,N}$, to be correlated with $(y_i)_{i=1,\ldots,N}$, where $N$ denotes the size of population, $U$; an example is the use of a proportion-to-size sampling design in the Current Employment Statistics (CES) survey of business establishments, administered by the U.S.
Bureau of Labor Statistics, for the purpose of measuring total employment by geographic area and industry. Higher unit inclusion probabilities are assigned to larger employers because they drive the variance of the resulting total employment estimator. Sampling designs that induce this correlation are termed “informative” and the balance of information in the sample is different from that in the population.

Savitsky & Toth (2016) proposed a plug-in estimator that formulates a sampling-weighted pseudo-posterior density by exponentiating each (last stage) unit-indexed likelihood contribution by a sampling weight constructed to be inversely proportional to the unit marginal inclusion probability, \( w_i \propto \frac{1}{\pi_i} \), where \( \pi_i = P(\delta_i = 1) \), for units, \( i = 1, \ldots, n \), where \( n \) denotes the number of units in the observed sample. An approximate, weight-exponentiated pseudo-likelihood for the population, \( \prod_{i=1}^{n} p(y_i|\lambda)^{\tilde{w}_i} \), if constructed from the \( n \) units observed in the sample.

The sampling weights, \( (w_i) \), are normalized, to \( (\tilde{w}_i) \), to control the amount of estimated posterior uncertainty. Savitsky & Toth (2016) default to normalizing, \( \sum_{i=1}^{n} \tilde{w}_i = n \). Novelo & Savitsky (2017) demonstrate that pseudo-posterior estimator constructed from weights normalized to \( n \) generally produce credibility intervals that fail to contract on frequentist confidence sets by under covering because they don’t account for dependencies among units induced by the joint distribution \( (P_{b_0}, P_\nu) \).

Our paper constructs a simple post-processing step that adjusts the scale and shape of sampling-weighted, pseudo-posterior parameter credibility sets that we show in the sequel achieves approximately correct coverage under a broad class of generally-used sampling designs. Our procedure applies an adjustment step to the posterior draws to achieve an asymptotic sandwich form for the pseudo-posterior covariance that is the same as that for the sampling-weighted pseudo-MLE. We accomplish the adjustment by computing the variance of the score function and the expectation of the square of its gradient under the
joint distribution, \((P_{\theta_0}, P_{\nu})\). The adjustment step is applied, numerically, by resampling the observed data, \(y_1, \ldots, y_n\), under an empirical distribution approximation for \((P_{\theta_0}, P_{\nu})\). Performing a re-sampling step reduced to simply drawing from the existing sample at those stages where dependence is limited within blocks of units. All units nested within each re-sampled block are included in each re-sample; for example, if the multistage design includes a clustering step, we use the known cluster memberships of the last stage units and just re-sample the clusters. The population generating distribution, \(P_{\theta_0}\), is estimated, once, on our original sample and the adjustment is evaluated using the best available estimate for \(\theta\), the posterior mean. Our adjustment is, therefore, computationally fast and achieves nearly correct coverage for \(\theta\). The pseudo-posterior MCMC sampler requires only a simple edit to the population posterior sampler (to insert weights) because the same posterior geometry is employed. Our adjustment procedure requires no change to the MCMC sampler for the pseudo-posterior, which preserves its ease of use.

Rao & Wu (2010) address the under coverage of pseudo-posterior in the specific case of formulating \(P_{g(Y)}\) as an empirical likelihood for the purpose of estimating a total or mean, \(\hat{g}(Y)\). They replace \(n\) as the normalizer for \((\tilde{w}_i)\) with \(n^* = n/\text{DEFF}_{g(Y)}\), where \(\text{DEFF}_{g(Y)} = \text{Var}_{P_{\nu}}(g(Y))/\text{Var}_{\text{SRS}}(g(Y))\) denotes the design effect, defined as the variance induced under sampling design distribution, \(P_{\nu}\), divided by that under simple random sampling (SRS). Their approach improves the coverage properties for estimation of simple statistics, rather than some \(\theta\) of interest to the data analyst for a general \(P_{\theta}\). In addition, the simultaneous modeling of multiple outcomes or parameters would require multiple DEFF’s to be used, which is not possible if DEFF is only incorporated via the scaling the sample size \(n\).

Ribatet et al. (2012) motivate a similar sandwich form of an adjustment of the asymptotic covariance of the pseudo-posterior distribution as we do we under specification of
a composite weight-exponentiated pseudo-likelihood, where their pseudo-likelihood is employed to approximate a likelihood that is not able to be specified. By contrast, our survey sampling set-up assumes existence of a population model, $P_{\theta_0}$, which, though unknown, has a tractable form that allows consistency of our estimator, $P_\theta$. Even though consistency is achieved, the survey weighted pseudo-posterior is still misspecified because the exponentially weighted likelihood is a noisy approximation to the true likelihood of the joint distribution $(P_{\theta_0}, P_\nu)$. The sampling-weighted pseudo-posterior arises out of a random sampling mechanism to approximate the information in the population using a partially observed sample taken from that population. So Ribatet et al. (2012) don’t compute expectations with respect to the joint distribution, $(P_{\theta_0}, P_\nu)$, to develop their adjustment since they do not contemplate a random sampling process governed by $P_\nu$. Ribatet et al. (2012) implicitly assume that their weight-adjusted pseudo-posterior is correct in the absence of an ability to specify the exact likelihood. By contrast, we provide theoretical results for the form of the asymptotic sampling-weighted pseudo-MLE covariance matrix under the joint distribution for population generation and the taking of a sample. Ribatet et al. (2012) re-design the MCMC sampler to accomplish the adjustment, unlike our post-processing step, such that their approach requires the development of a specialized MCMC sampler, distinct from the sampler developed for the population, $P_{\theta_0}$.

Novelo & Savitsky(2017) develop an alternative approach to the pseudo-posterior distribution that multiplicatively adjusts the likelihood to accomplish asymptotically unbiased estimation of the population model on the observed informative sample. Their adjustment specifies a conditional population model, $p(\pi_i|y_i)$, unlike the plug-in approach that treats inclusion probabilities as fixed under the pseudo-posterior formulation. They show that credible intervals estimated from their adjusted, fully Bayes posterior achieves correct coverage in the case of a simple, single stage proportion to size sampling design. Their
likelihood adjustment, however, requires a different MCMC sampler than that developed for the population model and the adjusted likelihood includes an integration that must be numerically computed in each MCMC draw. So the fully Bayesian estimator lacks the ease-of-implementation of the pseudo-posterior distribution. Our post-processing adjustment step applied to the pseudo-posterior MCMC samples corrects the under coverage demonstrated by Novelo & Savitsky (2017). The fully Bayes approach will tend to be produce more efficient credible sets, however, under the requirement to specify a conditional population model for the inclusion probabilities that is assumed to be consistent. In practice, sample designs are often algorithmically defined, becoming quite complex. The fully Bayes approach has not been applied to multistage cluster designs. The impact of clustering on the effective sample size may still be a challenge. In this work, we demonstrate that the survey-weighted pseudo-posterior can be adjusted to give correct inference even under complex survey designs which include within-cluster dependence.

1.1 Motivating Multistage Cluster Design: The National Survey on Drug Use and Health

Our motivating survey design is the National Survey on Drug Use and Health (NSDUH), sponsored by the Substance Abuse and Mental Health Services Administration (SAMHSA). NSDUH is the primary source for statistical information on illicit drug use, alcohol use, substance use disorders (SUDs), mental health issues, and their co-occurrence for the civilian, non institutionalized population of the United States. The NSDUH employs a multi-stage state-based design (Morton et al. 2016), with the earlier stages defined by geography within each state in order to select households (and group quarters) nested within these geographically-defined primary sampling units (PSUs). Williams & Savitsky (2018) pro-
vides conditions for asymptotic consistency for the pseudo-posterior for designs like the NSDUH, which are characterized by:

- Cluster sampling, such as selecting only one unit per cluster, or selecting multiple individuals from a dwelling unit.
- Population information such as socio-economic indicators used to sort sampling units along gradients.

Both features are common, in practice, and create sampling dependencies that do not attenuate even if the population grows. For simplicity of exposition we examine the relationship between two measures, current (past month) smoking of cigarettes and past year major depressive episode for adults through a two-parameter logistic regression model.

## 2 Asymptotic Covariance Matrix of the pseudo-posterior Distribution

### 2.1 Setup

We suppose random variables of the population are generated, \( \mathbf{X}_\nu = (X_1, \ldots, X_{N_\nu}) \overset{\text{ind}}{\sim} P_{\theta_0} \) where \( \theta_0 \in \mathbb{R}^d \) and we perform inference on \( \theta \in \Theta \) of the population model from the sample of size, \( n_\nu \). A sampling design imposes a known distribution on a vector of inclusion indicators, \( \delta_\nu = (\delta_{\nu 1}, \ldots, \delta_{\nu N_\nu}) \), on units composing a population, \( U_\nu \). The sampling distribution takes an observed random sample, \( S_\nu \subseteq U_\nu \), of size \( n_\nu \leq N_\nu \) from \( U_\nu \). Our conditions for the main results are based on 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 \( (\delta_{\nu 1}, \ldots, \delta_{\nu N_\nu}) \). We denote
the sampling distribution by \( P_\nu \), which governs the taking of samples from the population. \( P_\nu \) is implicitly conditionally defined given realizations from \( P_{\theta_0} \). In other words, the joint distribution for \( \delta_\nu \) can depend on the some population information from \( X_\nu \).

The inclusion probabilities are formulated to depend on the finite population data values, \( X_\nu \), so that we employ the pseudo-posterior estimator to approximate the population likelihood from the observed sample with,

\[
p^\pi (X_{\nu i} \delta_{\nu i}) := p (X_{\nu i})^{\frac{\delta_{\nu i}}{\pi_{\nu i}}}, \ i \in U_\nu, \tag{1}
\]
which weights each density contribution, \( p(X_{\nu i}) \), by the inverse of its marginal inclusion probability \([\text{Savitsky & Toth} 2016]\). This approximation for the population likelihood produces the associated pseudo-posterior density,

\[
p^\pi (\theta \mid X_\nu \delta_\nu) = \frac{\prod_{i \in U_\nu} p^\pi_\theta (X_{\nu i} \delta_{\nu i}) \pi(\theta)}{\int_\Theta \prod_{i \in U_\nu} p^\pi_\theta (X_{\nu i} \delta_{\nu i}) \pi(\theta) d\theta}, \tag{2}
\]

where \( X_\nu \delta_\nu = (X_{\nu 1} \delta_{\nu 1}, \ldots, X_{\nu N_\nu} \delta_{\nu N_\nu}) \) denotes the observed sample of size, \( n_\nu \). The pseudo-posterior mass placed on subset \( B \subseteq \Theta \) becomes

\[
\Pi^\pi (B \mid X_\nu \delta_\nu) = \int_{\theta \in B} \pi^\pi (\theta \mid X_\nu \delta_\nu) \pi(\theta) d\theta \tag{3}
\]

In typical applications \([\text{Savitsky & Srivastava} 2018]\), sampling weights are normalized to satisfy \( \sum_{i \in S_\nu} \pi_{\nu i}^{-1} = n_\nu \), which regulates the scale of uncertainty in the estimated pseudo-posterior distribution. In practice, dependencies induced by informative, multistage sampling designs produce a smaller effective sample size than \( n_\nu \), such that the typical procedure under-estimates posterior uncertainty. In addition, the shape (geometry) of the pseudo-posterior distribution is impacted by the dependence induced in each stage of the sampling design such that the asymptotic covariance matrix will not be the same as that for the MLE obtained under simple random sampling. We proceed to derive the form
of the limiting covariance matrix for the pseudo-MLE under informative sampling, which we define as the MLE of Equation 1. We demonstrate that the covariance matrix of the pseudo-MLE is different from that for the MLE under simple random sampling, but that the latter is a special case of the former. We next demonstrate that the limiting covariance matrix of the pseudo-posterior distribution differs from the pseudo-MLE under informative sampling (due to the failure of Bartlett’s second identity) such that resulting credibility intervals would not be expected to contract on valid frequentist confidence intervals, absent adjustment. The difference between the limiting covariance matrix for the pseudo-posterior distribution, on the one hand, from that for the MLE under simple random sampling, on the other hand, may only be partly driven by informativeness of the sampling design. The dependencies induced under employment of a multistage sampling design, such as the within cluster dependence of units, will also impact the scale and shape of the limiting covariance matrix of the pseudo-posterior distribution, even absent sampling informativeness. In other words, even where sampling inclusion probabilities, \((\pi_\nu_i)\), are not required to provide unbiased estimation of \(\theta \in \Theta\), the resulting limiting covariance matrix of the posterior distribution under multistage sampling would be different from that for the MLE under simple random sampling.

Our main result is achieved in the limit as \(\nu \uparrow \infty\), under the countable set of successively larger-sized populations, \(\{U_\nu\}_{\nu \in \mathbb{Z}^+}\). The asymptotics under our construction is controlled by \(\nu \in \mathbb{N}\) to map to the process where we fix a \(\nu\), construct an associated finite population of size, \(N_\nu\), generate random variables \(X_{\nu 1}, \ldots, X_{\nu N_\nu} \overset{\text{ind}}{\sim} P_{\theta_0}\), construct unit marginal sample inclusion probabilities, \((\pi_{\nu 1}, \ldots, \pi_{\nu N_\nu})\) under \(P_\nu\) and then draw a sample, \(\{1, \ldots, n_\nu\}\) from that population. The process is repeated for each increment of \(\nu\). We define the associated stochastic rates of convergences notations, \(A_\nu = o_P(B_\nu)\) to denote that \(A_\nu = Y_\nu B_\nu\) where \(Y_\nu \overset{P}{\rightarrow} 0\). and \(A_\nu = O_P(B_\nu)\) denotes \(A_\nu = Y_\nu B_\nu\) where \(Y_\nu = O_P(1)\). For deterministic
sequences, $A_\nu$ and $B_\nu$, the notations reduce to the usual $o$ and $O$.

2.2 Preliminaries

We will construct asymptotic distributions for the sequence of centered and scaled random quantities,

$$h_{N_\nu} = \sqrt{N_\nu} (\theta - \theta_0),$$

for specific estimators. Let $\hat{\theta}_{\pi,N_\nu}$ denote the MLE of the pseudo-likelihood in Equation 1 (that we denote as the pseudo-MLE), which defines the sequence,

$$\hat{h}_{\pi,N_\nu} = \sqrt{N_\nu} \left( \hat{\theta}_{\pi,N_\nu} - \theta_0 \right),$$

as contrasted with centered and scaled sequence for the MLE under simple random sampling (SRS),

$$\hat{h}_{N_\nu} = \sqrt{N_\nu} \left( \hat{\theta}_{N_\nu} - \theta_0 \right).$$

Define the log-likelihood, $\ell_\theta = \log p_\theta = \log p_{\theta_0+h_{N_\nu}/\sqrt{N_\nu}}$ and the associated score function, $\ell_\theta = \nabla_\theta \ell_\theta$.

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

$$P_{\pi,N_\nu} = \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} \frac{\delta_{\nu i}}{\pi_{\nu i}} \delta (X_{\nu i}),$$

where $\delta (X_{\nu i})$ denotes the Dirac delta function, with probability mass 1 on $X_{\nu i}$ and we recall that $N_\nu = |U_\nu|$ denotes the size of the finite population. This construction contrasts with the usual empirical distribution, $P_{N_\nu} = \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} \delta (X_{\nu i})$. 

11
We follow the notational convention of Ghosal et al. (2000) and define the associated expectation functionals with respect to these empirical distributions by $P_{\pi_N} f = \frac{1}{N} \sum_{i=1}^{N} \delta_{X_{\nu_i}} f (X_{\nu_i})$. Similarly, $P_{N} f = \frac{1}{N} \sum_{i=1}^{N} f (X_{\nu_i})$. Lastly, we use the associated centered empirical processes, $G_{\pi_N} = \sqrt{N} \left( P_{\pi_N} - P_0 \right)$ and $G_{N} = \sqrt{N} \left( P_{N} - P_0 \right)$.

We construct two variance expressions, starting with Fisher’s information:

$$H_{\theta_0} = -\frac{1}{N} \sum_{i \in U_{\nu}} \mathbb{E}_{P_{\theta_0}} \hat{\ell}_{\theta_0} (X_{\nu_i}),$$

whose inverse provides the asymptotic covariance of the pseudo-posterior under our Bernstein Von-Mises result that follows. Next, we define:

$$J_{\theta_0} = \frac{1}{N} \sum_{i \in U_{\nu}} \mathbb{E}_{P_{\theta_0}} \hat{\ell}_{\theta_0} (X_{\nu_i}) \hat{\ell}_{\theta_0} (X_{\nu_i})^T,$$

which is the middle term in the asymptotic variance of the MLE under simple random sampling. Under the population model (and an SRS sample), the likelihood is properly specified, so $J_{\theta_0} = H_{\theta_0}$.

Because our pseudo-posterior framework arises from a random sampling process governed by $P_{\nu}$,

$$H_{\theta_0}^\pi = -\mathbb{E}_{P_{\theta_0}} \left[ \mathbb{E}_{P_{\nu}} \left[ \frac{\delta_{X_{\nu_i}}}{P_{\nu}} \hat{\ell}_{\theta_0} \right] \right]_{\pi_{\nu_i}}$$

$$= -\frac{1}{N} \sum_{i \in U_{\nu}} \mathbb{E}_{P_{\theta_0}} \left[ \mathbb{E}_{P_{\nu}} \left[ \frac{\delta_{X_{\nu_i}} A_{\nu}}{P_{\nu}} \hat{\ell}_{\theta_0} \right] \right]$$

$$= -\frac{1}{N} \sum_{i \in U_{\nu}} \mathbb{E}_{P_{\theta_0}} \hat{\ell}_{\theta_0} (X_{\nu_i})$$

$$= H_{\theta_0},$$

where $A_{\nu}$ denotes the sigma field of information in $U_{\nu}$. This equivalence between $H_{\theta_0}^\pi$ and $H_{\theta_0}$ does not hold for the weighted composite likelihood of Ribatet et al. (2012), where the
weights are arbitrary and arise from a deterministic process to approximate an intractable likelihood for the population, $U_\nu$.

Our main results in the following section are anchored in the observation that the survey weighted $J_{\theta_0}^\pi = \mathbb{E}_{P_{\theta_0}, \nu} \left[ \mathbb{P}_{\nu, \ell_{\theta_0}} \hat{\ell}_{\theta_0} \right] \neq J_{\theta_0}$ due to the mis-specification from using a noisy approximation to the likelihood for $(P_{\theta_0}, P_\nu)$.

### 2.3 Main Results

The following conditions guarantee three results on the forms for asymptotic covariance matrices of the distributions for pseudo-MLE estimator and the pseudo-posterior. The first theorem extends Theorem 5.23 of van der Vaart (1998) to derive the asymptotic expansion of the centered and scaled pseudo-MLE. The second theorem specifies the form of the associated sandwich covariance matrix for the (asymptotic expansion of the) pseudo-MLE. The third theorem extends similar theorems in Kleijn & van der Vaart (2012) and van der Vaart (1998) that specify the covariance matrix of the asymptotic Gaussian form for the pseudo-posterior distribution. We observe that the asymptotic covariance matrices are different for each of the MLE, the pseudo-MLE and the pseudo-posterior, which sets up our proposed scale and shape adjustment, introduced in the sequel.

**(A1) (Continuity)** For each $\theta \in \Theta \in \mathbb{R}^d$ (an open subset of Euclidean space), $\ell_{\theta_0}(x)$ be a measurable function (of $x$) and differentiable at $\theta_0$ for $P_{\theta_0}$—almost every $x$ (with derivative, $\dot{\ell}_{\theta_0}(x)$), such that for every $\theta_1$ and $\theta_2$ in a neighborhood of $\theta_0$ with $\mathbb{E}_{\theta} \dot{\ell}_{\theta_0}(x) \dot{\ell}_{\theta_0}(x)^T < \infty$, we have a Lipschitz condition:

$$\left| \ell_{\theta_1}(x) - \ell_{\theta_2}(x) \right| \leq \dot{\ell}_{\theta_0}(x) \| \theta_1 - \theta_2 \| \text{a.s. } P_{\theta_0}$$

**(A2) (Local Quadratic Expansion)** The Kullback-Liebler divergence with respect to $P_{\theta_0}$
has a second order Taylor expansion about \( \theta_0 \),

\[
\mathbb{E}_{p_{\theta_0}} \log \frac{p_\theta}{p_{\theta_0}} = \frac{1}{2} (\theta - \theta_0)^T H_{\theta_0} (\theta - \theta_0) + o\left(\|\theta - \theta_0\|^2\right),
\]

where \( H_{\theta_0} \) is a \( d \times d \) positive definite matrix.

(A3) (Bartlett’s First Identity)

\[\mathbb{E}_{p_{\theta_0}} \dot{\ell}_{\theta_0} = 0\]

(A4) (Consistency of the MLE for the population)

\[\mathbb{P}_{N_\nu} \ell_{\theta_{N_\nu}} \geq \sup_\theta \mathbb{P}_{N_\nu} \ell_\theta - \mathcal{O}_{P_{\theta_0}} \left( N_\nu^{-1} \right)\]

and \( \hat{\theta}_{N_\nu} \xrightarrow{P_{\theta_0}} \theta_0 \)

(A5) (Non-zero Inclusion Probabilities)

\[
\sup_\nu \left[ \frac{1}{\min_{i \in U_\nu} |\pi_{\nu i}|} \right] \leq \gamma, \text{ with } P_{\theta_0} - \text{probability 1.}
\]

(A6) (Growth of dependence is restricted)

For every \( U_\nu \) there exists a binary partition \( \{S_{\nu 1}, S_{\nu 2}\} \) of the set of all pairs \( S_\nu = \{\{i, j\} : i \neq j \in U_\nu\} \) such that

\[
\limsup_{\nu \uparrow \infty} |S_{\nu 1}| = \mathcal{O}(N_\nu),
\]

and

\[
\limsup_{\nu \uparrow \infty} \max_{i,j \in S_{\nu 2}} \left| \frac{\pi_{\nu ij}}{\pi_{\nu i} \pi_{\nu j}} - 1 \right| = \mathcal{O}(N_\nu^{-1}), \text{ with } P_{\theta_0} - \text{probability 1}
\]
We note that Conditions \( \text{(A4) - (A6)} \) are necessary to produce consistency of the sampling weighted pseudo-posterior estimator and, by extension the MLE of the sampling weighted likelihood,

\[
P_{\pi,N}^\ell_{\theta,\pi,\nu} \geq \sup_{\theta,\nu} P_{\pi}^\ell_{\theta} - \sigma_{P_{\theta_0},P_{\nu}} \left( N_\nu^{-1} \right)
\]

and \( \hat{\theta}_{\pi,\nu} \overset{P_{\theta_0},P_{\nu}}{\rightarrow} \theta_0 \). See Savitsky & Toth (2016), Williams & Savitsky (2018) for more details.

**Theorem 1.** Suppose conditions \( \text{(A1)-(A6)} \) hold. Then

\[
\sqrt{N_\nu} \left( \hat{\theta}_{\pi,\nu} - \theta_0 \right) = -H_{\theta_0}^{-1} \frac{1}{\sqrt{N_\nu}} \sum_{i=1}^{N_\nu} \frac{\delta_{\pi,\nu i}}{\pi_{\nu i}} \hat{\ell}_{\theta_0}(X_{\nu i}) + \sigma_{P_{\theta_0},P_{\nu}}(1) \tag{10}
\]

\[
= -H_{\theta_0}^{-1} \sqrt{N_\nu} \mathbb{E}_{P_{\nu}} \hat{\ell}_{\theta_0} + \sigma_{P_{\theta_0},P_{\nu}}(1) \tag{11}
\]

\[
= -H_{\theta_0}^{-1} \mathbb{E}_{P_{\nu}} \hat{\ell}_{\theta_0} + \sigma_{P_{\theta_0},P_{\nu}}(1). \tag{12}
\]

**Theorem 2.** Suppose conditions \( \text{(A1)-(A5)} \) hold. Then

\[
\text{Var}_{P_{\theta_0},P_{\nu}} - H_{\theta_0}^{-1} \sqrt{N_\nu} \mathbb{E}_{P_{\nu}} \hat{\ell}_{\theta_0} = H_{\theta_0}^{-1} J_{\theta_0}^\pi H_{\theta_0}^{-1} \tag{13a}
\]

\[
= H_{\theta_0}^{-1} \left[ J_{\theta_0} + \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} \mathbb{E}_{P_{\theta_0}} \left\{ \left[ \frac{1}{\pi_{\nu i}} - 1 \right] \hat{\ell}_{\theta_0}(X_{\nu i})\hat{\ell}_{\theta_0}(X_{\nu i})^T \right\} \right] H_{\theta_0}^{-1} \tag{13b}
\]

\[
\leq \gamma H_{\theta_0}^{-1} J_{\theta_0} H_{\theta_0}^{-1} = \gamma H_{\theta_0}^{-1}. \tag{13c}
\]

The upper bound in Equation \( \text{(13c)} \) demonstrates that \( \gamma \) injures the \( \sqrt{N_\nu} \) convergence rate achieved for the MLE (under simple random sample of size, \( N_\nu \), the population size) for the convergence of the pseudo-MLE. The larger is \( \gamma \), the more varied will be information in the samples around that for the population, which indicates a decreasing efficiency of the sampling design. The amount of injury would be higher for less efficient sampling
designs. The maximum penalty paid is a uniformly inflated scale, which will produce wider confidence regions. Even in the case of simple random sampling of sample size, \( n_\nu < N_\nu \), however, while the inclusion probabilities will be equal, the value will be less than 1, so that the convergence rate is injured from \( \sqrt{N_\nu} \) to \( \sqrt{n_\nu} \), which is the same as the rate of the MLE under simple random sampling of size \( n_\nu \). Thus the SRS design is a special case. Equation \[13b\] demonstrates that the shape or geometry of the limiting distribution will be impacted in the case of unequal sampling inclusion probabilities. This “warping” effect would be expected to be more pronounced in a highly-skewed proportion-to-size sampling design than in an unequally-weighted stratified sampling design with relatively few strata. The upper bound in Theorem 2 does not restrict the possibility that some designs may be more efficient than an SRS or that the efficiency varies by parameter. We demonstrate both of these phenomena via simulations in Section 4.

**Theorem 3.** Suppose conditions (A1)-(A6) hold. Then

\[
\sup_{B \in \Theta} \left| \Pi_{N_\nu}^\chi (\theta \in B \mid X_\nu, \delta_\nu) - N_{\hat{\theta}_{\pi,N_\nu}, N_\nu}^{-1} H_{\theta_0}^{-1} (B) \right| P_{\theta_0}^{P_\nu} \to 0, \quad (14)
\]

where \( \hat{\theta}_{\pi,N_\nu} \) may be the pseudo-MLE or the pseudo-posterior mean.

The different forms of the asymptotic covariance matrices for the pseudo-MLE, on the one hand, and the pseudo-posterior, on the other hand, are driven by the failure of Bartlett’s second identity under informative sampling. This difference motivates our post-processing step, which we next introduce, that performs multiplicative adjustments to draws from the pseudo-posterior distribution such that their covariance is approximately equal to that of the pseudo-MLE.
3 Post-processing the pseudo-posterior

From Section 2 we see that the asymptotic covariance of the pseudo-MLE or pseudo-posterior mean is $H_{\theta_0}^{-1} J_\theta H_{\theta_0}^{-1}$, yet the asymptotic covariance of our samples drawn from the pseudo-posterior is $H_{\theta_0}^{-1}$. This is analogous to the differences observed in Ribatet et al. (2012), though our formulation for $J_\theta$ (and also $H_{\theta_0}$) arises from a random-sampling mechanism, which we leverage in the sequel to perform a post-hoc adjustment to draws from the pseudo-posterior. Let $\hat{\theta}_m$ represent the sample from the pseudo-posterior for $m = 1, \ldots, M$ draws with sample mean $\bar{\theta}$. Define the adjusted sample:

$$\hat{\theta}_m^a = \left(\hat{\theta}_m - \bar{\theta}\right) R_2^{-1} R_1 + \bar{\theta},$$

(15)

where $R_1^\prime R_1 = H_{\theta_0}^{-1} J_\theta H_{\theta_0}^{-1}$ and $R_2^\prime R_2 = H_{\theta_0}^{-1}$. We may loosely think of $R_2^{-1} R_1$ as a multivariate ‘design effect’ adjustment (For the SRS sample, we expect Barlett’s second identity to hold and thus $H_{\theta_0}^{-1} J_\theta H_{\theta_0}^{-1} = H_{\theta_0}^{-1}$ which is the same asymptotic variance as the unadjusted pseudo-posterior). Since $\hat{\theta}_m \sim N(\theta_0, N\nu^{-1} H_{\theta_0}^{-1})$, we now have $\hat{\theta}_m^a \sim N(\theta_0, N\nu^{-1} H_{\theta_0}^{-1} J_\theta H_{\theta_0}^{-1})$, which is the asymptotic distribution of the MLE under the pseudo-likelihood. Unlike Ribatet et al. (2012), who pre-compute the MLE and change the geometry of their posterior sampler, our implementation is applied as a post-hoc projection of the pseudo-posterior sample, leaving the initial Monte Carlo sampler intact. So the data analyst may use the Monte Carlo sampler that they designed for population model estimation (under simple random sampling).

For composite likelihoods, Ribatet et al. (2012) calculate $\text{Var}_{P_\theta_0} \hat{\theta}_0 = J_\theta$ analytically. However, we have an additional distribution $P_\nu$ for the sampling design which is unlikely to be in analytic form. In practice, the design is often algorithmically defined, for example designs may use the sorting and clustering of population units in addition to unequal probabilities of selection. Rather than assuming a simplifying model for this distribution,
we instead approximate the joint distribution \((P_{\theta_0}, P_{\nu})\) with the empirical distribution by resampling the units and associated response values.

Under a multi-stage sampling design with primary sampling units (PSUs) constructed as blocks (e.g., geographic regions or households) of last stage units (e.g., persons), we would re-sample a subset of the PSUs that contain dependent last-stage units, followed by including all last-stage units with each PSU. We use information about the PSU memberships of each last stage unit in the observed sample in order to conduct the resampling. The data analyst is expected to have this information about the structure of the sampling design, in addition to possessing the sampling weights for the last stage units (e.g., persons). It is necessary when conducting the resampling to explicitly re-sample blocks of units, such as PSUs, when member units express dependence. Such a procedure preserves the dependence structure within the replicate re-samples. This resampling procedure ensures our adjustment properly estimates the scale inflation of the pseudo posterior distribution induced by the dependent step(s). We use a simple random sampling without replacement (SRSWOR) procedure to re-sample the PSUs because they are nearly independent from one another, in practice.

Algorithm 1 provides a simple and computationally efficient resampling approach to estimate of \(\text{Var}_{P_{\theta_0}, P_{\nu}} \left[ \mathbb{P}_{N_{\nu}} \hat{\ell}_{\theta_0} \right] = J_{\theta_0}^\pi\). We recall from Section 2.2 that \(H_{\theta_0} = -\mathbb{E}_{P_{\theta_0}} \hat{\ell}_{\theta_0}\) and \(H_{\pi}^\theta = -\mathbb{E}_{P_{\theta_0}, P_{\nu}} \left[ \mathbb{P}_{N_{\nu}} \hat{\ell}_{\theta_0} \right] = H_{\theta_0}\). Therefore, consistent estimates of \(H_{\theta_0}\) are available without Algorithm 1. Both the plug-in estimate \(-\sum_{i \in S} w_i \hat{\ell}_{\hat{\theta}_m}(X_i)\) and the posterior average \(-\frac{1}{M} \sum_{m=1}^{M} \sum_{i \in S} w_i \hat{\ell}_{\theta_m}(X_i)\) using the original sample \(S\) will provide consistent estimates of \(H_{\theta_0}\). (We drop the “\(\nu\)” subscript from \(X\) for readability). However estimating \(\hat{H}_{\theta_0}\) via Algorithm 1 is convenient during the estimation of \(\hat{J}_{\theta_0}^\pi\), which cannot be estimated without replication. For simplicity, we use half the PSUs from the sample in each replicate. Other resampling without replacement approaches should be effective. However, sampling the
PSUs with replacement under-estimates the variance when the number of PSUs nested within strata is very small because with replacement sampling inaccurately reproduces the sampling design of PSUs from the population. For example, the NSDUH sample only has two PSUs available per strata.
Algorithm 1: Adjust pseudo-posterior to correct for complex survey design

input: \( \hat{\theta}_m \) from the pseudo-posterior \( \{2\} \)
{\( j, k \)} indicators for PSUs \( j = 1, \ldots, J_k \) and Strata \( k = 1, \ldots, K \).
{\( w_{ijk}, X_{ijk} \)} for all \( i \) in \( 1, \ldots, I_{jk} \) for every \( \{j, k\} \)
\( R \) number of replicates.

output: Adjusted sample \( \hat{\theta}_m^{a} \)

1. Calculate the posterior mean \( \bar{\theta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m \)

2. for Replicas \( r \leftarrow 1 \) to \( R \) do

   | Subsample PSUs without replacement (SRSWOR) |
   |---------------------------------------------|
   | for Strata \( k \leftarrow 1 \) to \( K \) do |
   | Sample half the PSUs within strata \( k \): \( \{j^\prime\}_k^r \) with \( |\{j^\prime\}_k| = J_k/2 \) |
   | Take all units within each selected PSU: \( S^r_k = \cup_{j^\prime} \cup \{ij^\prime k\} \) |
   | Define sample \( \{w^r_l, X^r_l\} \) for new index \( l \in S^r_k \) |
   | Double weights \( \hat{w}^r_l = 2w^r_l \) |

3. end

4. Combine samples across strata: \( S^r = \cup_k S^r_k \)

5. Normalize weights \( \hat{w}^r_l = \hat{w}^r_l \left( \frac{n}{\sum_{l \in S^r} \hat{w}^r_l} \right) \)

6. Evaluate \( h_r = -\sum_{l \in S^r} \hat{w}^r_l \ell_\theta(X^r_l) \) and \( j_r = \sum_{l \in S^r} \hat{w}^r_l \ell_\theta(X^r_l) \)

10. end

11. Calculate \( \hat{H}_\theta = \frac{1}{R} \sum_{r=1}^{R} h_r \)

12. Calculate \( \hat{J}_\theta = \frac{1}{R-1} \sum_{r=1}^{R} (j_r - \bar{j})(j_r - \bar{j})^t \) with \( \bar{j} = \frac{1}{R} \sum_{r=1}^{R} j_r \)

14. Calculate \( \hat{R}_1 \) via Cholesky decomposition: \( \hat{R}_1^t \hat{R}_1 = \hat{H}_\theta^{-1} \hat{J}_\theta \hat{H}_\theta^{-1} \)

15. Calculate \( \hat{R}_2 \) via Cholesky decomposition: \( \hat{R}_2^t \hat{R}_2 = \hat{H}_\theta^{-1} \)

17. Calculate inverse \( \hat{R}_2^{-1} \)

19. Evaluate Eq. 15: \( \hat{\theta}_m^{a} = \left( \hat{\theta}_m - \bar{\theta} \right) \hat{R}_2^{-1} \hat{R}_1 + \bar{\theta} \)

20.
4 Simulation Study

We construct a population model to address our inferential interest of a binary outcome \( y \) with a linear predictor \( \mu \).

\[
y_i \mid \mu_i \sim \text{Bern}(F_i^{-1}(\mu_i)), \ i = 1, \ldots, N
\]  

(16)

where \( F_i^{-1} \) is the quantile function (inverse cumulative function) for the logistic distribution. The first set of simulations (Section 4.1.1) is based on equal probability sampling. We let \( \mu \) depend on a single predictor \( x_1 \). The second set of simulations (Section 4.1.2) is based on unequal probability sampling. We let \( \mu \) depend on two predictors \( x_1 \) and \( x_2 \), where \( x_2 \) is a size variable to set the selection probabilities into the sample. The third set of simulations (Section 4.1.3) is also based on unequal probability sampling, but we let \( \mu \) depend on three predictors \( x_1, x_2, \) and \( z_2 \), where the latter is a random cluster effect at the PSU level. The quantity of inferential interest for all of our simulations is the estimation of the population model coefficients (intercept and slope) for \( x_1 \), and \( (x_2, z_2) \) are nuisance.

The variable \( x_1 \) represents the observed information available for analysis, whereas \( x_2 \) represents auxiliary information available for setting inclusion probabilities used to conduct sampling, which is either ignored or not available for analysis. The \( x_1 \) and \( x_2 \) distributions are \( \mathcal{N}(0, 1) \) and \( \mathcal{E}(r = 1/5) \) with rate \( r \), where \( \mathcal{N}(\cdot) \) and \( \mathcal{E}(\cdot) \) represent normal and exponential distributions, respectively. The cluster effect \( z_2 \) is neither a design variable used for sampling nor part of the analytical model, but is a nuisance representing unknown and un-modeled dependence between units within the same cluster (PSU). We choose \( z_2 \sim \mathcal{E}(1/5) \) for a skewed distribution.

We formulate the logarithm of the sampling-weighted pseudo-likelihood for estimating
where $p_i = F^{-1}_i(\mu_i)$ and the sampling weights, $w_i^*$ are normalized such that the sum of the weights equals the sample size $\sum_{i=1}^{n} w_i^* = n$.

Finally, we estimate the joint posterior distribution using Equation (17) coupled with our prior distribution assignments, using the NUTS Hamiltonian Monte Carlo algorithm implemented in Stan (Carpenter 2015, Stan Development Team 2016). All computations were performed in R (R Core Team 2017). Analytic functions for $\ell_\theta(X_i)$ and $\ddot{\ell}_\theta(X_i)$ were obtained by creating a function for $\ell_\theta(X_i)$ and then using the ‘deriv’ function in R to automatically generate functions for the gradient and hessian.

4.1 Simulation Designs

In the following subsections we discuss how we construct sampling design distributions, $P_\nu$, that will induce dependence and skewed information about the population in the observed sample as a means of assessing the performance of our post-processing adjustment procedure specified in Algorithm 1. In Section 4.2 we will assess whether the adjustments performed to the posterior draws generate credibility sets that achieve nominal frequentist coverage. We recall from Section 1 that the survey sampling literature defines the design effect (DEFF) as the ratio of the variance of a estimate for the population mean $\bar{Y}$ under a complex survey design compared to the variance under simple random sampling:

$$\text{DEFF}_{\bar{Y}} = \frac{\text{Var}_{P_\nu}(\hat{Y})}{\text{Var}_{SRS}(\hat{Y})}.$$ 

In addition to nominal coverage, we are also interested
in comparing our model-based design effects to the standard DEFF output of design-based survey software, such as the R ‘survey’ package [Lumley 2016]. We estimate the marginal design effect for each parameter: \( \text{DEFF}_\theta = \frac{\text{diag}\{H_\theta^{-1}J_\theta H_\theta^{-1}\}}{\text{diag}\{H_\theta^{-1}\}} \). These parameter-specific DEFFs provide an estimate of the marginal rescaling induced by the complex sample design relative to a simple random sample.

### 4.1.1 Equal Probability Dependent Designs (DE)

For these designs, we induce dependence in the observed samples by clustering units; for example, by aggregating individuals in the population by geographically-indexed domains. This type of clustering or grouping of units is performed by the sampling designers, in practice, in order to control the costs (in this case, travel and labor costs) of administering the survey. It is typically the case that the clustering structure will be coincident with a dependence structure in the population variables of interest; for example, geographically-indexed domains capture a spatial dependence among measures for individuals induced by similarities in culture and economic factors. The effect is that individuals are sampled in dependent groups or clusters, which is expected to lower the amount of information about the population in a realized random sample under this type of sampling design as compared to a simple random sampling of individuals taken from the same population. Even if a sampling design distribution, \( P_\nu \), is not informative, the design will induce a scale inflation in the asymptotic covariance of the posterior distribution if the design includes a stage that samples dependent clusters. Our theoretical results don’t directly address this possibility but instead focus on warping and scale adjustments due to approximation error of the pseudo posterior induced by unequal weighting. However, we demonstrate in the sequel that our post-processing adjustment procedure of Algorithm 1, nevertheless, adjusts the scale of the posterior distribution under this scenario to achieve nominal coverage.
The population generating model is

\[ \mu_i = 0.0 + 1.0 x_{1i} \]

where the intercept was chosen such that the median of \( \mu \) is 0, therefore the median of \( F_i^{-1}(\mu) \) is 0.5.

The first design (DE1), is a one-stage cluster design where clusters of size 5 are selected according to simple random sampling (SRS). All individuals have responses that are unconditionally independent. In other words, the clustering membership is randomized and uninformative. Under this scenario, the pseudo-likelihood reduces to the true likelihood with correctly specified independence between units. Therefore, both the unadjusted MCMC samples \( \hat{\theta}_m \) and the adjusted MCMC samples \( \hat{\theta}_n \) should ideally have similar coverage.

The second design (DE5) is the also a one-stage SRS design, except that all 5 members of each cluster have complete dependence. Both the \( y \) and the \( x_1 \) have identical values within each cluster: \( y_{ij} = y_{i'j} \) and \( x_{1ij} = x_{1i'j} \) for all individuals \( i \neq i' \) in cluster \( j \). Under this scenario, the pseudo-likelihood is again reduced to the simple likelihood. While the likelihood is correctly specified for any given individual, joint cluster dependence is misspecified as independence. Effectively, the sum of the (equal) weights should really sum to \( n/5 \) rather than \( n \). Under this scenario, the unadjusted MCMC samples \( \hat{\theta}_m \) should have intervals that are too narrow by a factor of \( \sqrt{5} \) while the adjusted intervals for \( \hat{\theta}_n \) should be longer and achieve the nominal coverage. This idealized example, in which within cluster dependence is both unspecified in the analyst’s model and complete, demonstrates the sensitivity of the posterior (and pseudo-posterior) to the mis-specification of the effective sample size \( n \) and the robustness of Algorithm 1 to correct for this.
4.1.2 One stage unequal probability designs (PPS1)

For these next designs, we have no dependence induced by the clustering of units. Instead, we use an informative design $P_\nu$ which uses information from the population to sample units with unequal probabilities of selection; for example, selecting larger businesses with higher probability than smaller businesses in the Current Employment Statistics (CES) survey, administrated by the U.S. Bureau of Labor Statistics (BLS). In practice, these designs control costs because large businesses contribute proportionately more to estimates for industry totals, such as total production or number of employees. Further refinements to the design, such as stratification of units into size classes, also create statistical efficiencies by reducing the possibility of extreme sample outcomes (such as selecting a sample composed entirely of small businesses). Our theoretical results directly address these informative designs which lead to warping and scale effects due to the approximation error of the pseudo-posterior induced by unequal weighting. We demonstrate that our post-processing adjustment via Algorithm 1 achieves nominal coverage under these informative sampling designs.

The population generating model is now

$$\mu_i = -1.88 + 1.0x_{1i} + 0.5x_{2i}$$

where the intercept was chosen such that the median of $\mu$ is approximately 0, therefore the median of $F_i^{-1}(\mu)$ is approximately 0.5. The size measure used for sample selection is $\tilde{x}_{2i} = x_{2i} - \min_i(x_{2i}) + 1$.

Even though the population response $y$ was simulated with $\mu = f(x_1, x_2)$, we estimate the marginal models at the population level for $\mu = f(x_1)$. This exclusion of $x_2$ is analogous to the situation in which an analyst does not have access to all the sample design information and ensures that our sampling design instantiates informativeness (where $y$ is
correlated with the selection variable, \( x_2 \), that defines inclusion probabilities). In particular, we estimate the models under informative design scenarios and compare the population fitted models, \( \mu = f(x_1) \), to those from the samples. The first unequally weighted design is a one-stage probability proportional to size design (PPS1), where probabilities of selection are proportional to the size measure \( \pi_i \propto \tilde{x}_{2i} \). For the same population we also create a stratified design (SPPS1). We add this additional design because stratification is expected to improve the efficiency of the sampling design as compared to SRS because it will - on average - produce samples that are more informationally representative of the population, such that \( \text{DEFF}_\theta \) may be less than 1. We demonstrate the our scale adjustment adapts to more efficient, as well as less efficient, sampling designs. The population is sorted by size measure \( \tilde{x}_2 \) and then placed into 10 strata. We then select \( n/10 \) units from each strata \( k \) with \( \pi_{ik} \propto \tilde{x}_{2ik} \).

4.1.3 Three stage unequal probability designs (PPS3)

The last set of designs combines feature of the first two sets. In practice, multi-stage designs such as the NSDUH first select geographic PSUs (such as states, counties, census tracts, etc) in proportion to a measure of population size. This provides both cost savings (collecting data in geographic clusters) and statistical efficiencies (higher population areas represent more of the population total), especially when combined with geographic-based stratification (e.g. by state). The final stages for multi-stage surveys are often the household and individual. The effect is that individuals within each PSU cluster may likely have outcome measures related to others in their household and geographic cluster. The within PSU dependence and the unequal probabilities of selection will induce both a scale inflation and a warping in the asymptotic covariance of the posterior distribution. Our theoretical results dont directly address the rescaling due to within cluster dependence; however, our
post-processing adjustment procedure of Algorithm 1, nevertheless, adjusts both the scale and shape of the posterior distribution under this scenario to achieve nominal coverage.

The population generating model is now

\[ \mu_{ij} = -1.88 + 1.0x_{1ij} + 0.25x_{2ij} + 0.25z_{2j} \]

where \( z_{2j} \sim \mathcal{E}(1/5) \) is the random effect for PSU \( j \). The median of \( \mu \) is still close to 0, and the median of \( F_i^{-1}(\mu) \) is still close to 0.5. The size measure used for sample selection is \( \tilde{x}_{2i} = x_{2i} - \min_i(x_{2i}) + 1 \). Compared to the population model for PPS1 and SPPS1, the relationship between \( y \) and the size variable \( x_2 \) is weaker (0.25 vs. 0.50). This is often the case for household surveys compared to establishment surveys, because the amount of information available to the sample designer is much greater for establishments than for households.

The next design is a three-stage PPS design (PPS3), analogous to a household survey in which a geographic area is selected as a PSU, followed by a household (HH) and an individual. We employ a simplified, but broadly representative, version of the design used for NSDUH where we first select the PSU based on the size \( \tilde{x}_2 \) aggregated up to the PSU level. We next select 5 out of 10 HHs within each PSU, where the HH’s are sorted based on an aggregate size measure from \( \tilde{x}_2 \) and sampled systematically (i.e. every other one along the rank sorted list). Finally, 1 of 3 individuals are selected within each household in proportion to the individual size measure \( \tilde{x}_2 \). The nested sampling within PSU, the systematic sampling of HHs, and the mutually exclusive sampling of individuals within HHs creates a sampling dependence that does not attenuate (i.e. factor). See Williams & Savitsky (2018) for a richer discussion of the sources of sampling dependence.

We include a PSU level random effect \( z_{2j} \) to allow for the possibility of un-modeled population level dependence that coincides with the sample design induced dependence and
together reduce the effective sample size. For example, geographic covariates such as state or census tract may be related to the outcome of interest, but like $x_2$ they are unavailable to the analyst of a public use file due to confidentiality protections. We expect the unadjusted MCMC sample $\hat{\theta}_m$ to undercover both due to the warping effect from unequal weighting and due to the over-estimation of the effective sample size from the nuisance PSU dependence. We expect the adjusted MCMC sample $\hat{\theta}^a_m$ to capture this dependence, leading to wider uncertainty intervals with closer to nominal coverage.

Lastly, we include a stratified version of the design (SPPS3) in which the aggregate size variable for the PSUs is used to sort the cluster into 10 strata, which are then sampled in a three stage design. Since the size variable $x_2$ has a weaker relationship with the outcome, the impact of stratification will be weaker for SPPS3 compared to SPPS1. This example is the closest to our motivating NSDUH design and provides insight into the performance of Algorithm 1 when resampling PSUs nested with strata.

4.2 Results

Table 1 provides a summary of results for 100 realizations for each of the 6 designs based on a target nominal coverage of 90%. Total sample sizes of $n = 200$ were used to explore performance for moderate sample sizes. Each realization was resampled $R = 100$ times to generate an adjustment via Algorithm 1. We consider coverage estimates from 85% to 95% to be reasonably close to the nominal 90% given the simulation noise from the 100 realizations. Marginal coverage is assessed from the two sided intervals from sample quantiles ($q_{5}, q_{95}$). For simplicity, joint coverage is assessed by comparing the Mahalanobis distance $(\hat{\theta} - \theta)'Var(\hat{\theta})(\hat{\theta} - \theta)$ to the 90% quantile of a $\chi^2_2$ distribution. Figure 1 displays one realization from each of the design simulations before and after adjustment to visually demonstrate the rescaling and rotation (to undo warping from unequal probability
informative sampling) of the adjustment.

### 4.2.1 Coverage

The one-stage equal probability designs (DE1 and DE5) demonstrate that Algorithm [1] is effective across the entire range of within cluster independence to complete cluster dependence. DE1 serves as a control, in which no adjustment should be needed. The marginal coverage and interval widths for the adjusted sample $\hat{\theta}_m^a$ are slightly lower than for the unadjusted $\hat{\theta}_m$ but the joint elliptical coverage is about as good. DE5 serves as an extreme example under which $\hat{\theta}_m$ is clearly undercovering and $\hat{\theta}_m^a$ performs much better. Figure [1] shows one realization in which the densities mostly overlap for DE1. For DE5, we see the adjusted density is much more diffused and indicates some design-induced dependence between the parameters. This may explain why the joint coverage for $\hat{\theta}_m$ is even worse than the marginal and it suggests that a naive rescaling of the weights by 5 might not lead to correct joint coverage as postulated in section [4.1.1].

The one-stage unequal probability designs (PPS1 and SPPS1) demonstrate the warping effect without the presence of within cluster dependence. PPS1 demonstrates improvements for both marginal and joint coverage. The stratified version (SPPS1) shows that the unadjusted sample $\hat{\theta}_m$ is over-covering, particularly for the joint region. For the moderate sample size $n = 200$, the adjusted coverage shows a decrease for the intercept but a much closer to nominal coverage for the joint region (88% vs. 99%). Figure [1] shows a similar pattern. The increase in dispersion for the PPS1 design is reduced and offset by stratification in SPPS1. These designs are similar to establishment surveys such as the CES, which may use frame data to form efficient strata and samples for businesses.

The three-stage designs with PSU level dependence (PPS3 and SPPS3) show similar results. The unequal selection is weaker in the three stage designs than in the one-stage.
Therefore the stratification does not lead to much gain in efficiency. Both designs show an improvement in coverage for both marginal and joint coverage. This is consistent with results from the one-stage designs, but combines the unequal weighting, within PSU dependence, and stratification into a single design, similar to household surveys such as the NSDUH.

4.2.2 Design Effects

We the compare the parameter-specific DEFF$\theta$ to the DEFF$\bar{y}$ based on Taylor linearization (Lumley 2004). Table 1 shows that the design effect for the intercept $\theta_0$ is very similar to the overall design effect for $y$, where the latter is computed from Lumley (2004). This is not surprising, since an intercept is very close to a mean. Examining the design effect for the slope $\theta_1$, we see that the effect of the design is typically less dramatic than for the intercept but still notably different from 1. We remind the reader that these estimates for design effects assume the bias has been removed due to incorporation of the weights and do not suggest that equally weighted likelihoods will lead to estimates for slopes that have correct coverage. For comparisons between consistent weighted estimates and biased unweighted estimates see Savitsky & Toth (2016), Williams & Savitsky (2018).
| Scenario | Marginal $\hat{\theta}_0$ | Marginal $\hat{\theta}_1$ | Joint $\hat{\theta}_0, \hat{\theta}_1$ | Width $\hat{\theta}_0$ | Width $\hat{\theta}_1$ | DEFF |
|----------|-----------------|-----------------|-----------------|-----------------|-----------------|------|
|          | $\hat{\theta}_m$ | $\hat{\theta}_m^a$ | $\hat{\theta}_m$ | $\hat{\theta}_m^a$ | $\hat{\theta}_m$ | $\hat{\theta}_m^a$ | $\hat{\theta}_m$ | $\hat{\theta}_m^a$ | $\hat{\theta}_m$ | $\hat{\theta}_m^a$ | $y$ |
| DE1      | 0.89            | 0.89            | 0.89            | 0.87            | 0.52            | 0.51            | 0.64            | 0.63            | 0.96            | 0.99            | 0.97 |
| DE5      | 0.43            | 0.56            | 0.89            | 0.87            | 0.52            | 1.24            | 0.70            | 1.60            | 5.06            | 5.26            | 5.10 |
| PPS1     | 0.77            | 0.83            | 0.83            | 0.91            | 0.50            | 0.69            | 0.55            | 0.70            | 1.90            | 1.63            | 1.86 |
| SPPS1    | 0.91            | 0.96            | 0.96            | 0.96            | 0.49            | 0.41            | 0.54            | 0.55            | 0.69            | 1.02            | 0.71 |
| PPS3     | 0.74            | 0.79            | 0.79            | 0.87            | 0.51            | 0.75            | 0.57            | 0.75            | 2.20            | 1.71            | 2.13 |
| SPPS3    | 0.77            | 0.80            | 0.80            | 0.87            | 0.51            | 0.73            | 0.56            | 0.71            | 2.03            | 1.59            | 1.99 |

Table 1: Summary of coverage, average width, and design effect estimates for simulations based on 90% posterior intervals. Based on $M = 100$ realizations with sample size $n = 200$, $R = 100$ replications, and population sizes $N = 5000, 5000, 6000$ for the SRS (DE1,DE5), one-stage PPS (PPS1,SPPS1), and and three-stage (PPS3, SPPS3) designs, respectively, where $S$ denotes the nesting within a stratified sampling stage.
Figure 1: MCMC pseudo-posterior sample for the intercept (horizontal) and slope (vertical) for one realization of each of six sample designs. Unadjusted (red circles) and adjusted (blue triangles) with approximate 90% density ellipses. Created with ‘ggplot2’ (Wickham 2009).
A simple logistic model relating current (past month) smoking status to the presence of a past year major depressive episode (MDE) was fit via the survey weighted pseudo-posterior as described in Section 4 using probability-based analysis weights for adults from the 2014 NSDUH public use data set. Overall the design effect for $y$ from the R ‘survey’ package is 1.87, and the parameter specific design effects based on the comparing interval widths for $\hat{\theta}_m^a$ and $\hat{\theta}_m$ are 1.88 for the intercept and 1.12 for the slope. In addition to the marginal rescaling, Figure 2 demonstrates the presence of a joint warping effect from the complex sample design of the NSDUH.

Estimates for $\theta$ agree to 4 decimal places when comparing the model fit using the R ‘survey’ package function ‘svyglm’ and the pseudo-posterior (-1.2817, 0.7130). The standard error estimates also closely match for both the intercept (0.0169 vs. 0.0170) and for the slope (0.0439 vs. 0.0433) when comparing the adjusted MCMC samples $\hat{\theta}_m^a$ to the pseudo-MLE estimates from ‘svyglm’. Given the large sample size of approximately 42,000 adults, this strong agreement is expected.
Figure 2: MCMC pseudo-posterior sample for the intercept (horizontal) and slope (vertical) for a logistic regression modeling current cigarette smoking by past year major depressive episode based the 2014 National Survey on Drug Use and Health. Unadjusted (red circles) and adjusted (blue triangles) draws with approximate 90% density ellipses. Created with ‘ggplot2’ (Wickham 2009).
6 Discussion

This work is motivated by the need to apply Bayesian models to survey data. Previous works (Savitsky & Toth 2016, Williams & Savitsky 2018) have demonstrated consistency of the survey weighted pseudo-posterior for a large class of population models and complex survey designs. However, Novelo & Savitsky (2017) observe that the resulting posterior intervals can have poor frequentist performance. Insights from the composite likelihood (Ribatet et al. 2012) and model mis-specification (Kleijn & van der Vaart 2012) literature motivated the development of the theory and adjustment of the asymptotic covariance of the survey weighted pseudo-posterior. This resulting adjusted pseudo-posterior can then be used for inference in the same manner as the posterior distribution from a simple random sample. It also achieves the same asymptotic properties as the pseudo-likelihood under ‘design-based’ frequentist inference methods. Therefore, these results allow for modelers to better incorporate informative sample design features into their analysis models while allowing survey statisticians to incorporate more complex modeling approaches into their analysis and production of official statistics.

Adjustment 15 implemented via Algorithm 1 provides a simple, computationally inexpensive, and effective approach to quantifying and adjusting for the warping of the pseudo-posterior due to unequal weighting and complex sampling dependence between sampling units. Our resampling algorithm eliminates the need to analytically integrate \( \text{Var}_{\hat{\theta}_0} \hat{\ell}_{\theta_0} \) and thus can be applied to the composite pseudo-likelihood as a more flexible alternative to the modified MCMC approaches presented in Ribatet et al. (2012). We note that adjustment 15 is a projection, but does not force the pseudo-posterior variance to equal that of the pseudo-MLE exactly for small-to-moderate samples. Instead, it provides an asymptotic adjustment which allows the analyst to base inference on the sample distribution of the
posterior (adjusted by the design effect) rather than using the asymptotic MLE covariance. If the latter is desired, benchmarking to force the posterior samples to exactly match specified mean and covariance can be performed in closed form using a constrained linear projection (Ghosh 1992, Datta et al. 2011) or via an iterative Newton-Raphson approach for other constrained projections (Williams & Berg 2013).

For the simplicity of exposition, the theory in Section 2 assumes the $X_i$ are conditionally independently drawn from $P_{\theta_0}$. Under this assumption, the unequal probability of selection from an informative sampling design induces warping of the pseudo-posterior covariance, as demonstrated in the PPS1 and SPPS1 simulations. However, within cluster dependence also reduces the effective sample size and may contribute to the warping, as is the case for simulations DE5, PPS3, and SPPS3 and the NSDUH example. As demonstrated by the simulation results, Algorithm 1 provides an adjustment to correct for both unequal probabilities of selection and within cluster dependence. However, the theory to incorporate un-modeled dependence for the population that coincides with sampling clusters is still incomplete. A more rigorous development of the theory for both modeled and un-modeled cluster dependence within a complex survey sample is needed. Consistency results have been established for a larger class of models than those considered here. Yet, even under SRS, Bernstein-von Mises results are much more complicated than consistency results, especially for non-parametric models (Ghosal & van der Vaart 2017). Therefore, extending the results of this work to combine complex survey design with more general classes of population models is also needed.
A Proof of Theorem 1

The proof strategy closely follows Theorem 5.23 of van der Vaart (1998) where we update the centered and scaled empirical process, $G_{N\nu}$, to its sampling-weighted extension, $G_{\pi N\nu}^\pi$. For every random sequence, $h_{N\nu}$, we extend van der Vaart (1998) Lemma 19.31 to achieve,

$$G_{\pi N\nu}^\pi \left( \sqrt{N\nu} \left( \ell_0 + \frac{h_{N\nu}}{\sqrt{N\nu}} - \ell_{\theta_0} \right) - h_{N\nu}^T \hat{\ell}_{\theta_0} \right) \xrightarrow{P_{\theta_0}, P_{\nu}} 0. \tag{18}$$

Conditions (A1) and (A3) along with

$$\mathbb{E}_{P_{\nu}} \left[ \mathbb{P}_{N\nu}^\pi, \ell_0 \right] = \mathbb{E}_{P_{\nu}} \left[ \frac{1}{N\nu} \sum_{i=1}^{N\nu} \delta_{\nu i} \ell_\theta(X_i) \right] \tag{19}$$

$$= \mathbb{P}_{N\nu} \ell_\theta \tag{20}$$

produces a 0 mean for the random sequence of Equation 18 with respect to the joint distribution, $(P_\theta, P_\nu)$. By the boundedness requirement for sequence $(\pi_{\nu i}^{-1})$ in Condition (A5), the Lipschitz condition in Condition (A1) and the dominated convergence theorem, their variance converges to 0 and the result in Equation 18 is achieved.

Conditions (A1), (A4), (A5) and Corollary 5.53 of van der Vaart (1998), the sequence $h_{N\nu} = \sqrt{N\nu} (\theta - \theta_0)$ is bounded in probability.

We may re-write Equation 18 as,

$$N\nu \mathbb{P}_{N\nu}^\pi \log \frac{p_{\theta_0 + \frac{h_{N\nu}}{\sqrt{N\nu}}}}{p_{\theta_0}} - h_{N\nu}^T G_{N\nu}^\pi \hat{\ell}_{\theta_0} = N\nu \mathbb{E}_{P_{\theta_0}} \log \frac{p_{\theta_0 + \frac{h_{N\nu}}{\sqrt{N\nu}}}}{p_{\theta_0}} = \sigma_{P_{\theta_0}, P_{\nu}} \tag{1}$$

From Condition (A2), we have,

$$\mathbb{E}_{P_{\theta_0}} \log \frac{p_{\theta_0 + \frac{h_{N\nu}}{\sqrt{N\nu}}}}{p_{\theta_0}} - \frac{1}{2N\nu} h_{N\nu}^T H_{\theta_0} h_{N\nu} = \sigma_{P_{\theta_0}} \tag{1}$$

Substituting this expression above yields,

$$N\nu \mathbb{P}_{N\nu}^\pi \log \frac{p_{\theta_0 + \frac{h_{N\nu}}{\sqrt{N\nu}}}}{p_{\theta_0}} = \frac{1}{2} h_{N\nu}^T H_{\theta_0} h_{N\nu} + h_{N\nu}^T G_{N\nu}^\pi \hat{\ell}_{\theta_0} + \sigma_{P_{\theta_0}, P_\nu} \tag{1}, \tag{21}$$

37
that we recognize as the local asymptotic normality condition of Kleijn & van der Vaart (2012) (which we will later use to derive the form for the asymptotic covariance matrix of the pseudo-posterior distribution). Equation 21 is true for both $\hat{h}_{N_\nu}$ and $\tilde{h}_{N_\nu} = - H_{\theta_0}^{-1} G_{N_\nu} \hat{\ell}_{\theta_0}$ by Condition (A4). The remainder of the proof exactly follows van der Vaart (1998) where we separately plug in each of $\hat{h}_{N_\nu}$ and $\tilde{h}_{N_\nu}$ for $h_{N_\nu}$ into Equation 21 to achieve two equivalent equations (up to $O_{\mathbb{P}}(1)$). We take the difference between the two equations and complete the square, which produces the result of the theorem.

B Proof of Theorem 2

We begin by constructively expanding the variance with respect to the joint distribution, $(P_{\theta_0}, P_\nu)$,

$$\text{Var}_{P_{\theta_0}, P_\nu} H_{\theta_0}^{-1} \sqrt{N_\nu} P_{N_\nu} \hat{\ell}_{\theta_0} = N_\nu H_{\theta_0}^{-1} \text{Var}_{P_{\theta_0}, P_\nu} P_{N_\nu} \hat{\ell}_{\theta_0} H_{\theta_0}^{-1}. \quad (22)$$

We proceed to apply the total variance decomposition to the variance of the random sequence in the middle of the above expression,

$$\text{Var}_{P_{\theta_0}, P_\nu} P_{N_\nu} \hat{\ell}_{\theta_0} = \text{Var}_{P_{\theta_0}} E_{P_\nu} \left[ P_{N_\nu} \hat{\ell}_{\theta_0} | A_\nu \right] + E_{P_{\theta_0}} \text{Var}_{P_\nu} \left[ P_{N_\nu} \hat{\ell}_{\theta_0} | A_\nu \right], \quad (23)$$

38
where, fixing $\nu$, $\mathcal{A}_\nu$ denotes the sigma field of information in the population, $U_\nu$. Next, we constructively evaluate each of the two terms.

$$
\begin{align*}
\text{Var}_{P_{\theta_0}} \mathbb{E}_{P_{\theta_0}} \left[ \mathbb{P}^\pi_{N_\nu} \hat{\ell}_{\theta_0} \mid \mathcal{A}_\nu \right] &= \text{Var}_{P_{\theta_0}} \left\{ \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} \mathbb{E}_{\pi_{\nu_i}} [\delta_{\nu_i} \mid \mathcal{A}_\nu] \hat{\ell}_{\theta_0} (X_i) \right\} \\
&= \text{Var}_{P_{\theta_0}} \left\{ \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} \hat{\ell}_{\theta_0} (X_i) \right\} \\
&= \frac{1}{N_\nu^2} \mathbb{E}_{P_{\theta_0}} \left\{ \sum_{i=1}^{N_\nu} \hat{\ell}_{\theta_0} (X_i) \right\}^2 \\
&= \frac{1}{N_\nu^2} \left[ \sum_{i=1}^{N_\nu} \mathbb{E}_{P_{\theta_0}} \hat{\ell}_{\theta_0} (X_i) \hat{\ell}_{\theta_0} (X_i)^T + \sum_{i \neq j \in U_\nu} \mathbb{E}_{P_{\theta_0}} \hat{\ell}_{\theta_0} (X_i) \hat{\ell}_{\theta_0} (X_j)^T \right] \\
&= \frac{1}{N_\nu^2} \sum_{i=1}^{N_\nu} \mathbb{E}_{P_{\theta_0}} \hat{\ell}_{\theta_0} (X_i) \hat{\ell}_{\theta_0} (X_i)^T,
\end{align*}
$$

where the second term in the second equation from the bottom results because $X_i \perp X_j$ under $P_{\theta_0}$ and by Condition (A3).
\[ \mathbb{E}_{P_{\theta_0}} \text{Var}_{P_{\nu}} \left[ \mathbb{P}_{N_{\nu}} \hat{\ell}_{\theta_0} \mid A_{\nu} \right] = \mathbb{E}_{P_{\theta_0}} \left\{ \frac{1}{N^2_{\nu}} \text{Var}_{P_{\nu}} \left[ \sum_{i=1}^{N_{\nu}} \frac{\delta_{\nu i}}{\pi_{\nu i}} \hat{\ell}_{\theta_0} \right] \mid A_{\nu} \right\} \]  

\[ = \frac{1}{N^2_{\nu}} \mathbb{E}_{P_{\theta_0}} \left\{ \sum_{i=1}^{N_{\nu}} \text{Var}_{P_{\nu}} \left[ \frac{\delta_{\nu i}}{\pi_{\nu i}} \hat{\ell}_{\theta_0} \right] \hat{\ell}_{\theta_0} (X_i) \hat{\ell}_{\theta_0} (X_i)^T \right\} \]  

\[ + \sum_{i \neq j \in U_{\nu}} \text{Cov}_{P_{\nu}} \left[ \frac{\delta_{\nu i} \delta_{\nu j}}{\pi_{\nu i} \pi_{\nu j}} \right] \hat{\ell}_{\theta_0} (X_i) \hat{\ell}_{\theta_0} (X_j)^T \right\} \]  

\[ = \frac{1}{N^2_{\nu}} \sum_{i=1}^{N_{\nu}} \mathbb{E}_{P_{\theta_0}} \left\{ \left[ \frac{1}{\pi_{\nu i}} - 1 \right] \hat{\ell}_{\theta_0} (X_i) \hat{\ell}_{\theta_0} (X_i)^T \right\} \]  

\[ + \frac{1}{N^2_{\nu}} \sum_{i \neq j \in U_{\nu}} \mathbb{E}_{P_{\theta_0}} \left\{ \left[ \frac{\pi_{\nu i j}}{\pi_{\nu i} \pi_{\nu j}} - 1 \right] \hat{\ell}_{\theta_0} (X_i) \hat{\ell}_{\theta_0} (X_j)^T \right\} \]  

\[ \leq \frac{1}{N^2_{\nu}} \sum_{i=1}^{N_{\nu}} \mathbb{E}_{P_{\theta_0}} \left\{ \left[ \frac{1}{\pi_{\nu i}} - 1 \right] \hat{\ell}_{\theta_0} (X_i) \hat{\ell}_{\theta_0} (X_i)^T \right\} \]  

\[ + \max \{1, \gamma - 1\} \frac{1}{N^2_{\nu}} \sum_{i \neq j \in U_{\nu}} \mathbb{E}_{P_{\theta_0}} \left\{ \hat{\ell}_{\theta_0} (X_i) \hat{\ell}_{\theta_0} (X_j)^T \right\} \]  

\[ = \frac{1}{N^2_{\nu}} \sum_{i=1}^{N_{\nu}} \mathbb{E}_{P_{\theta_0}} \left\{ \left[ \frac{1}{\pi_{\nu i}} - 1 \right] \hat{\ell}_{\theta_0} (X_i) \hat{\ell}_{\theta_0} (X_i)^T \right\} . \]  

The sequence, \(-1 \leq \left\{ \frac{\pi_{\nu i j}}{\pi_{\nu i} \pi_{\nu j}} - 1 \right\}\), in Equation 25c is bounded from above by \( \left\{ \frac{1}{\pi_{\nu i}} - 1 \right\} \leq (\gamma - 1) \) by Condition [A5]. See Williams & Savitsky (2018) for more details. The second expression in Equation 25d exactly equals 0 by the independence of \( X_i \) and \( X_j \) (\( \forall i \neq j \in U_{\nu} \)) under \( P_{\theta_0} \) and by Condition [A3]. Since the second expression in Equation 25d is bounded from above by 0, it exactly equals 0 (for all \( \nu \in \mathbb{Z}^+ \), \( i \neq j \in U_{\nu} \)), producing the
equality in Equation 25c. Equation 25c results from the following computations:

\[
\text{Var}_{P_\nu} \left[ \frac{\delta_{\nu i}}{\pi_{\nu i}} \mid A_\nu \right] = \mathbb{E}_{P_\nu} \left[ \frac{\delta_{\nu i}}{\pi_{\nu i}} \mid A_\nu \right]^2 - \left[ \mathbb{E}_{P_\nu} \frac{\delta_{\nu i}}{\pi_{\nu i}} \mid A_\nu \right]^2 \\
= \frac{1}{\pi_{\nu i}} - 1
\]

\[
\text{Cov}_{P_\nu} \left[ \frac{\delta_{\nu i} \delta_{\nu j}}{\pi_{\nu i} \pi_{\nu j}} \mid A_\nu \right] = \mathbb{E}_{P_\nu} \left[ \frac{\delta_{\nu i} \delta_{\nu j}}{\pi_{\nu i} \pi_{\nu j}} \mid A_\nu \right] - \mathbb{E}_{P_\nu} \left[ \frac{\delta_{\nu i}}{\pi_{\nu i}} \mid A_\nu \right] \mathbb{E}_{P_\nu} \left[ \frac{\delta_{\nu j}}{\pi_{\nu j}} \mid A_\nu \right] \\
= \frac{\pi_{\nu ij}}{\pi_{\nu i} \pi_{\nu j}} - 1
\]

We plug in the results for Equations 24 and 25 back into Equation 23.

\[
N_\nu \text{Var}_{P_0, P_\nu} \mathbb{E}_{N_\nu} \hat{\ell}_{\theta_0} = \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} \mathbb{E}_{P_0} \hat{\ell}_{\theta_0}(X_i) \hat{\ell}_{\theta_0}(X_i)^T + \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} \mathbb{E}_{P_0} \left\{ \left[ \frac{1}{\pi_{\nu i}} - 1 \right] \hat{\ell}_{\theta_0}(X_i) \hat{\ell}_{\theta_0}(X_i)^T \right\} \\
\leq \gamma J_{\theta_0}
\]

and the result is achieved.

### C Proof of Theorem 3

The proof strategy is the same as Kleijn & van der Vaart (2012) where they first prove the assertion on any two sets of random sequences, \((h_{N_\nu}, g_{N_\nu}) \in K\), where \(K \in \mathbb{R}^d\) is an arbitrary compact set. They then extend the result to a sequence of balls, \(K_{N_\nu}\), centered on 0 with increasing radii, \(M_{N_\nu} \uparrow \infty\). We extend their strategy by updating notation to incorporate the \((\delta_{\nu i}, \pi_{\nu i})\), where \(\pi_{\nu i} = \text{Pr}\{\delta_{\nu i} = 1\}\), governed by the sampling design distribution, \(P_\nu\), such that our result applies for \((P_0, P_\nu)\), jointly. Recall that we have the
local asymptotic normality result,
\[ N_{\nu} \mathbb{P}_{N_{\nu}} \log \frac{P_{\theta_0} + \frac{h_{N_{\nu}}}{\sqrt{N_{\nu}}}}{P_{\theta_0}} = \frac{1}{2} h_{N_{\nu}}^T H_{\theta_0} h_{N_{\nu}} + h_{N_{\nu}}^T \frac{\partial}{\partial \theta_0} \dot{\ell}_{\theta_0} + \sigma_{P_{\theta_0}, P_{\nu}}(1), \]  

from the proof of Theorem 1 under Conditions (A1), (A2), (A3) and (A5). Define the sampling weighted empirical loglikelihood ratio,
\[ s_{N_{\nu}}(h) = N_{\nu} \mathbb{P}_{N_{\nu}} \log \frac{P_{\theta_0} + \frac{h_{N_{\nu}}}{\sqrt{N_{\nu}}}}{P_{\theta_0}}, \]

and let \( \Delta_{N_{\nu}, \theta_0} = H_{\theta_0}^{-1} \frac{\partial}{\partial \theta_0} \dot{\ell}_{\theta_0} \). Plugging into Equation 27, we achieve,
\[ s_{N_{\nu}}(h_{N_{\nu}}) = h_{N_{\nu}}^T H_{\theta_0} \Delta_{N_{\nu}, \theta_0} - \frac{1}{2} h_{N_{\nu}}^T H_{\theta_0} h_{N_{\nu}} + \sigma_{P_{\theta_0}, P_{\nu}}(1). \]

Let \( \phi_{N_{\nu}} \) denote the normal distribution, \( \mathcal{N}(\Delta_{N_{\nu}, \theta_0}, H_{\theta_0}^{-1}) \) and define the sequence of random functions,
\[ f_{N_{\nu}}(g_{N_{\nu}}, h_{N_{\nu}}) = \left( 1 - \frac{\phi_{N_{\nu}}(h_{N_{\nu}}) s_{N_{\nu}}(g_{N_{\nu}}) \pi_{N_{\nu}}(g_{N_{\nu}})}{\phi_{N_{\nu}}(g_{N_{\nu}}) s_{N_{\nu}}(h_{N_{\nu}}) \pi_{N_{\nu}}(h_{N_{\nu}})} \right). \]

Plugging into the logarithm of Equation 29 for \( s_{N_{\nu}}(\cdot) \) and \( \phi_{N_{\nu}}(\cdot) \), where for any \((h_{N_{\nu}}, g_{N_{\nu}}) \in K\), the prior ratio, \( \pi_{N_{\nu}}(g_{N_{\nu}})/\pi_{N_{\nu}}(h_{N_{\nu}}) \to 1 \) as \( \nu \uparrow \infty \), we achieve:
\[ \log \frac{\phi_{N_{\nu}}(h_{N_{\nu}}) s_{N_{\nu}}(g_{N_{\nu}}) \pi_{N_{\nu}}(g_{N_{\nu}})}{\phi_{N_{\nu}}(g_{N_{\nu}}) s_{N_{\nu}}(h_{N_{\nu}}) \pi_{N_{\nu}}(h_{N_{\nu}})} = (g_{N_{\nu}} - h_{N_{\nu}})^T H_{\theta_0} \Delta_{N_{\nu}, \theta_0} + \frac{1}{2} h_{N_{\nu}}^T H_{\theta_0} h_{N_{\nu}} - \frac{1}{2} g_{N_{\nu}}^T H_{\theta_0} g_{N_{\nu}} + \sigma_{P_{\theta_0}, P_{\nu}}(1) \]
\[ - \frac{1}{2} (g_{N_{\nu}} - \Delta_{N_{\nu}, \theta_0})^T H_{\theta_0} (h_{N_{\nu}} - \Delta_{N_{\nu}, \theta_0}) + \frac{1}{2} (g_{N_{\nu}} - \Delta_{N_{\nu}, \theta_0})^T H_{\theta_0} (g_{N_{\nu}} - \Delta_{N_{\nu}, \theta_0}) \]
\[ = \sigma_{P_{\theta_0}, P_{\nu}}(1), \]

as \( \nu \uparrow \infty \). Conclude that
\[ \sup_{g, h \in K} f_{N_{\nu}}(g, h) \xrightarrow{P_{\theta_0}, P_{\nu}} 0, \]
as $\nu \uparrow \infty$. Define $\Xi_{N_\nu}$ as the event that $\Pi_{N_\nu}^\pi(K) > 0$. Define $\Pi_{N_\nu}^\pi,K(B \mid X_\nu^\delta_\nu) = \Pi_{N_\nu}^\pi(h \in B \mid X_\nu^\delta_\nu) / \Pi_{N_\nu}^\pi(K \mid X_\nu^\delta_\nu)$ to the posterior mass truncated to the compact space, $K$, and similarly for $\Phi_{N_\nu}^K$. Fix (any) $\eta > 0$ and define the sequence of events, $\Omega_{N_\nu} = \{\sup_{g,h \in K} f_{N_\nu}^\pi(g,h) \leq \eta\}$. Construct the inequality,

$$
\mathbb{E}_{P_{0\theta},P_\nu} \left\| \Pi_{N_\nu}^\pi,K - \Phi_{N_\nu}^K \right\|_1 \mathbb{I}_{\Xi_{N_\nu}} \leq \mathbb{E}_{P_{0\theta},P_\nu} \left\| \Pi_{N_\nu}^\pi,K - \Phi_{N_\nu}^K \right\|_1 \mathbb{I}_{\Omega_{N_\nu} \cap \Xi_{N_\nu}} + 2 \mathbb{E}_{P_{0\theta},P_\nu}(\Xi_{N_\nu} \setminus \Omega_{N_\nu}),
$$

(32)

where the total variation normal, $\|\cdot\|$, is bounded above by 2 and the second on the right-hand side is $O(1)$ from Equation 31. Since $\|P - Q\| = 2 \int (1 - p/q)^+ dQ$, we may expand the first term on the right-hand side,

$$
\frac{1}{2} \mathbb{E}_{P_{0\theta},P_\nu} \left\| \Pi_{N_\nu}^\pi,K - \Phi_{N_\nu}^K \right\|_1 \mathbb{I}_{\Omega_{N_\nu} \cap \Xi_{N_\nu}} \leq \mathbb{E}_{P_{0\theta},P_\nu} \int \left(1 - \frac{\phi_{N_\nu}(h)s_{N_\nu}^\pi(g)\pi_{N_\nu}(g)}{\phi_{N_\nu}(g)s_{N_\nu}^\pi(h)\pi_{N_\nu}(h)}\right) d\Phi_{N_\nu}^K(g) d\Pi_{N_\nu}^\pi,K(h) \mathbb{I}_{\Omega_{N_\nu} \cap \Xi_{N_\nu}} \leq \eta.
$$

(33) (34)

(35)

The proof next follows Kleijn & van der Vaart (2012) to expand the result on a compact $K$ to compact sets, $(K_{N_\nu})_{\nu}$ of balls with radii $M_{N_\nu} \uparrow \infty$, which provides the result for $\mathbb{R}^d$ in the limit of $\nu$. From Theorem 1 we have:

$$
\hat{h}_{\pi,N_\nu} = \sqrt{N_\nu} \left(\hat{\theta}_{\pi,N_\nu} - \theta_0\right) = -\Delta_{N_\nu,\theta_0}^\pi + o_{P_{0\theta},P_\nu}(1),
$$

(36)

and the stated result is achieved with a rescaling and shift since the total variation norm is invariant to rescalings and shifts.

References

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
of Statistics 34(1), 86–102.

URL: http://EconPapers.repec.org/RePEc:bla:scjsta:v:34:y:2007:i:1:p:86-102

Carpenter, B. (2015), ‘Stan: A probabilistic programming language’, Journal of Statistical Software.

Datta, G. S., Ghosh, M., Steorts, R. & Maples, J. (2011), ‘Bayesian benchmarking with applications to small area estimation’, TEST 20(3), 574–588.

URL: https://doi.org/10.1007/s11749-010-0218-y

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. (2017), Fundamentals of Nonparametric Bayesian Inference, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Ghosh, M. (1992), ‘Constrained bayes estimation with applications’, Journal of the American Statistical Association 87(418), 533–540.

URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1992.10475236

Kleijn, B. & van der Vaart, A. (2012), ‘The bernstein-von-mises theorem under misspecification’, Electron. J. Statist. 6, 354–381.

URL: https://doi.org/10.1214/12-EJS675

Lumley, T. (2004), ‘Analysis of complex survey samples’, Journal of Statistical Software 9(1), 1–19. R package version 2.2.

Lumley, T. (2016), ‘survey: analysis of complex survey samples’. R package version 3.32.
Morton, K. B., Aldworth, J., Hirsch, E. L., Martin, P. C. & Shook-Sa, B. E. (2016), Section 2, sample design report, in ‘2014 National Survey on Drug Use and Health: Methodological Resource Book’, Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration, Rockville, MD.

Novelo, L. L. & Savitsky, T. (2017), ‘Fully Bayesian Estimation Under Informative Sampling’, ArXiv e-prints.

URL: https://arxiv.org/abs/1710.00019

R Core Team (2017), R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria.

URL: https://www.R-project.org/

Rao, J. N. K. & Wu, C. F. J. (2010), ‘Bayesian pseudo-empirical-likelihood intervals for complex surveys’, Journal of the Royal Statistical Society - Series B 72, 533–544.

Ribatet, M., Cooley, D. & Davison, A. C. (2012), ‘Bayesian inference from composite likelihoods, with an application to spatial extremes’, Statistica Sinica 22(2), 813–845.

URL: http://www.jstor.org/stable/24310036

Savitsky, T. D. & Srivastava, S. (2018), ‘Scalable bayes under informative sampling’, Scandinavian Journal of Statistics pp. n/a–n/a. 10.1111/sjos.12312.

URL: http://dx.doi.org/10.1111/sjos.12312

Savitsky, T. D. & Toth, D. (2016), ‘Bayesian Estimation Under Informative Sampling’, Electronic Journal of Statistics 10(1), 1677–1708.

Stan Development Team (2016), ‘RStan: the R interface to Stan’. R package version 2.14.1.

URL: http://mc-stan.org/
van der Vaart, A. W. (1998), *Asymptotic statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Wickham, H. (2009), *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.

**URL:** [http://ggplot2.org](http://ggplot2.org)

Williams, M. & Berg, E. (2013), ‘Incorporating user input into optimal constraining procedures for survey estimates’, *Journal of Official Statistics* 29(3), 375 – 396.

Williams, M. R. & Savitsky, T. D. (2018), ‘Bayesian Estimation Under Informative Sampling with Unattenuated Dependence’, *ArXiv e-prints*.

**URL:** [https://arxiv.org/abs/1807.05066](https://arxiv.org/abs/1807.05066)