Aust. N. Z. J. Stat. 2019 doi: 10.1111/j.1467-842X.XXX

Bayesian weighted inference from surveys

David Gunawan[1], Anastasios Panagiotelis[2], William Griffiths[3]*and Duangkamon Chotikapanich[2]

University of Wollongong, Monash University and University of Melbourne

Summary

Data from large surveys are often supplemented with sampling weights that are designed to reflect unequal probabilities of response and selection inherent in complex survey sampling methods. We propose two methods for Bayesian estimation of parametric models in a setting where the survey data and the weights are available, but where information on how the weights were constructed is unavailable. The first approach is to simply replace the likelihood with the pseudo likelihood in the formulation of Bayes theorem. This is proven to lead to a consistent estimator but also leads to credible intervals that suffer from systematic undercoverage. Our second approach involves using the weights to generate a representative sample which is integrated with a Markov chain Monte Carlo (MCMC) or other simulation algorithm designed to estimate the parameters of the model. In extensive simulation studies, the latter methodology is shown to achieve performance comparable to the standard frequentist solution of pseudo maximum likelihood, with the added advantage of being applicable to models that require inference via MCMC. The methodology is demonstrated further by fitting a mixture of gamma densities to a sample of Australian household income.

Key words: sampling weights; latent representative sample; Markov chain Monte Carlo; gamma mixture; pseudo maximum likelihood.

1. Introduction

Raw data from surveys seldom come from a simple random sample where selection of each individual is equiprobable, but instead from complex survey sampling methods such as stratification and multistage sampling that exhibit unequal probabilities of selection and non-response. Examples of large surveys with these characteristics are the Panel Study of Income Dynamics (PSID), the British Household Panel Survey (BHPS), and the Household Income and Labour Dynamics in Australia (HILDA) survey, all of which are increasingly

*Author to whom correspondence should be addressed.
[1] School of Mathematics and Applied Statistics, University of Wollongong, NSW, 2522, Australia
[2] Department of Econometrics and Business Statistics, Monash University, Victoria, 3800, Australia
[3] Department of Economics, University of Melbourne, Victoria, 3010, Australia
Email: wegrif@unimelb.edu.au
Acknowledgment. The authors acknowledge the comments of anonymous referees that have greatly improved the paper.

used in applied statistical research. For samples that are non-representative in the sense that individuals with different characteristics have different probabilities of selection, the standard methods of inference and estimation may be biased or inconsistent; this issue is discussed in detail by Korinek, Mistiaen & Ravallion (2007), Pfeffermann (1996), Breunig (2001), and Wooldridge (1999, 2001, 2007).

A common way to address the problem is to use sampling weights provided with the survey datasets. Sampling weights act as expansion factors that scale and correct the representativeness of the sample to the population. They accommodate complex sampling designs and may be modified to ensure demographics such as sex, race, and age from the weighted sample match known census figures. If a survey respondent comes from a demographic group that has a low probability of selection or response, they are allocated a higher weight. Because sampling weights must take into account a large number of factors, their computation is often complicated (see Gelman et al. 2013; Korinek, Mistiaen & Ravallion 2007, and references therein), and detailed information on how they were constructed may not be available to researchers. We are concerned with a situation typical in much applied work where the only available information is the dataset and the sampling weights for each unit in the sample, with little or no information regarding the complex sampling design or how the weights were computed. In line with this information set, we treat the sampling weights as given and do not focus on their estimation and construction.

A number of methodologies that exploit survey weights to obtain unbiased and consistent estimation and inference have been proposed. One of the earliest approaches for estimating the population mean of a random variable is the classical weighted ratio estimator (see Horvitz & Thompson 1952). The most popular framework for taking sampling weights into account when estimating parametric models is pseudo maximum likelihood; see, for example, Godambe & Thompson (1986), Molina & Skinner (1992), Hesketh & Skrondal (2006), Skinner & Mason (2012), and references therein. To obtain the pseudo maximum likelihood estimator (PMLE), the usual log-likelihood is replaced with an objective function that is the sum of each sample weight multiplied by the contribution of its corresponding observation to the log likelihood. The resulting estimator is a special case of a general inverse probability weighted M-estimator (Wooldridge 1999, 2001, 2007).

There have also been a number of papers tackling the issue of survey weights from a Bayesian perspective. Aitkin (2008) and Rao & Wu (2010) incorporate sampling weights into pseudo Bayesian methods for a multinomial empirical likelihood, leading to Dirichlet posterior distributions. They provide Bayesian interval estimates for the population mean that are asymptotically valid in a frequentist framework. Using poststratification of cells based on sampling weights, Si, Pillai & Gelman (2015) developed a multinomial model for cell counts and a Bayesian nonparametric regression model for modelling an outcome variable conditional on the weights. More recently, Savitsky & Toth (2016) considered a Bayesian pseudo posterior, which is proportional to the product of the pseudo likelihood and a prior distribution. Adopting a nonparametric approach, where the true data generating process (DGP) is an unknown distribution within a density space, they prove that the pseudo posterior is a consistent estimator of the DGP.

In contrast to these earlier studies, we are concerned with accounting for sampling weights using Bayesian inference for the parameters in a parametric model and, as well as consistency, we are also concerned with precision as reflected by frequentist coverage in repeated samples. A procedure along these lines is useful if the fundamental aim is to base inferences on the posterior distributions of parameters, and quantities of interest that are functions of those parameters. It is also useful for exploiting numerical methods such as Markov chain Monte Carlo (MCMC) for estimating complex statistical models that are handled more easily within a Bayesian rather than a likelihood framework, such as mixtures, or multinomial and multivariate probit models.

We consider two approaches for incorporating the information from sampling weights into Bayesian inference. The first, which we call the Bayesian Pseudo Posterior Estimator (BPPE) simply replaces the likelihood with the pseudo-likelihood in the usual formulation of Bayes theorem. This is the approach taken by Savitsky & Toth (2016), but they are concerned with consistent estimation of an unknown density; we are concerned with inference for the parameters of a potentially complex model. The second approach, which we call the Bayesian Weighted Estimator (BWE), is a data-augmentation approach where a pseudo representative sample is treated as missing data. We consider two approaches for generating a pseudo representative sample; the first is resampling with replacement from the observed data using the normalized sampling weights, while the second is an algorithm from Dong, Elliott & Raghunathan (2014a), based on the weighted finite population Bayesian Bootstrap. Inference about the unknown parameters can be conducted via MCMC as if the pseudo representative sample were the data. Since the early work of Tanner & Wong (1987), data augmentation has been used extensively for Bayesian estimation of a variety of statistical models. See, for example, Chib (1992), Albert & Chib (1993), Geweke & Keane (2007) and Geweke & Amisano (2011).

Replacing the likelihood with some other function of the parameters and data is an idea that goes at least as far back as the notion of proper likelihoods introduced by Monahan & Boos (1992) and has received significant treatment in the case of Bayesian empirical likelihood (see Lazar 1989; Schennach 2005; Rao & Wu 2010). We evaluate the asymptotic behavior of our two proposed approaches under an assumption of non-informative priors. For the BPPE we are able to derive theoretical results that suggest consistency, but an asymptotic variance that leads to undercoverage of credible intervals in repeated sampling.

These theoretical results are validated in a simulation study. In the case of the BWE, the likelihood is replaced with a Monte Carlo estimate of a density that is a discrete mixture over all possible pseudo-samples. Although this mixture is difficult to work with theoretically, we provide a sound intuitive justification for its use, and show through extensive simulations that the Bayesian weighted estimators that we propose can achieve accurate empirical coverage.

We begin Section 2 with a brief description of the PMLE and its sandwich covariance matrix estimator, followed by a discussion of the problems that arise if this approach is adopted within a Bayesian framework. The details of our proposal for an alternative Bayesian weighted estimator that utilises generation of a representative sample are presented in Section 3. In Section 4 we use two simulation studies to illustrate application of the proposed estimator and to compare its repeated sampling properties to those of alternative estimators. Two quite different models are chosen for these illustrations: estimation of the mean and variance of a Gaussian distribution, and estimation of the parameters of a two-component mixture of gamma densities. In Section 5 Bayesian weighted and unweighted estimates of an Australian income distribution, modelled as a three component mixture of gamma densities, are presented. A conclusion is provided in Section 6.

2. Pseudo likelihood approaches

Assume we have a random variable Y whose population can be described by the density function $p(Y|\boldsymbol{\theta})$, $\boldsymbol{\theta}$ being an unknown vector of parameters we wish to estimate. We are supplied with a non-representative sample $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ that is based on a complex survey design, typically involving several demographic factors. Corresponding to each sample observation, we are also supplied with sampling weights $\boldsymbol{w} = (w_1, \ldots, w_n)^\top$, $0 < w_i < \infty$, but the details of the survey design and how the weights are calculated are not available to the investigator. It is assumed that the weights have been constructed such that a weight w_i is inversely proportional to the probability that the survey design selected an observation with the demographic characteristics of observation y_i. For estimation, observations whose probability of being selected is less than it would be under simple random sampling are weighted more heavily than they would be under simple random sampling, and vice versa. We assume that the w_i have been scaled such that $\sum_{i=1}^n w_i = n$. In what follows we first briefly describe the pseudo maximum likelihood estimator for $\boldsymbol{\theta}$ (Section 2.1), followed by a Bayesian estimator that uses the pseudo likelihood function (Section 2.2). Our proposal for a Bayesian weighted estimator designed to overcome problems with using the pseudo likelihood within a Bayesian framework is described in Section 3.

2.1. Pseudo maximum likelihood estimator

A pseudo log likelihood is defined as $L_p(\boldsymbol{\theta};\boldsymbol{y}) = \sum_{i=1}^{n} w_i \log p(y_i|\boldsymbol{\theta})$. The PMLE $\hat{\boldsymbol{\theta}}_{PML}$ satisfies the first order conditions

$$\frac{\partial L_p(\boldsymbol{\theta};\boldsymbol{y})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} w_i \frac{\partial \log p(y_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0},$$

This estimator is consistent but not efficient (Wooldridge 1999, 2001, 2007). Under some regularity conditions $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{PML} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(\mathbf{0}, \boldsymbol{H}_w^{-1}\boldsymbol{V}_w\boldsymbol{H}_w^{-1}\right)$, where $\boldsymbol{\theta}_0$ is the true value for $\boldsymbol{\theta}$ and $\boldsymbol{H}_w$ and $\boldsymbol{V}_w$ are consistently estimated using

$$\hat{\boldsymbol{H}}_w = \frac{1}{n}\sum_{i=1}^{n} w_i \left.\frac{\partial^2 \log p(y_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{PML}},$$

and

$$\hat{\boldsymbol{V}}_w = \frac{1}{n}\sum_{i=1}^{n} w_i^2 \left.\frac{\partial \log p(y_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \log p(y_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{PML}},$$

respectively. For making inferences about $\boldsymbol{\theta}$ the standard errors are obtained from the observed sandwich covariance estimator $n^{-1}\hat{\boldsymbol{H}}_w^{-1}\hat{\boldsymbol{V}}_w\hat{\boldsymbol{H}}_w^{-1}$ (White 1980, 1982).

2.2. Bayesian pseudo posterior estimator

Given the successful development of the pseudo likelihood sampling theory approach to estimating $\boldsymbol{\theta}$, a natural question to ask is whether a Bayesian approach with the usual likelihood function replaced by the pseudo likelihood would be suitable. For a given prior distribution $p(\boldsymbol{\theta})$, the posterior density obtained using this approach is given by

$$\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{w}) \propto p(\boldsymbol{\theta})\prod_{i=1}^{n} p(y_i|\boldsymbol{\theta})^{w_i}.$$

Theorem 1. Asymptotic properties of pseudo-posterior. *The pseudo posterior $\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{w})$ converges to a normal distribution with mean $\hat{\boldsymbol{\theta}}$ and covariance matrix $-n\hat{\boldsymbol{H}}_w^{-1}$ where $\hat{\boldsymbol{\theta}}$ is the posterior mode and $-n\hat{\boldsymbol{H}}_w^{-1} = n^{-1}\sum_{i=1}^{n} w_i \partial^2 \log p(y_i|\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top$ is the weighted Hessian.*

Corollary 1. *The posterior mode $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}_0$ where $\boldsymbol{\theta}_0$ is a unique solution to the population maximisation problem $\boldsymbol{\theta}_0 = \max_{\boldsymbol{\theta}_0\in\boldsymbol{\Theta}} \mathrm{E}_Y\left[\log p(Y|\boldsymbol{\theta})\right]$.*

Proof. See Appendix.

Since the pseudo posterior distribution converges to a normal distribution with a covariance matrix which differs from that of the PMLE, interval estimates derived from it will not have the correct frequentist coverage, a property usually regarded as desirable, even for Bayesian estimators. This is apendixpparent in our Monte Carlo simulations where these intervals suffer from undercoverage of the true parameter. Another disadvantage of this approach is that simple algorithms based on conjugate, or at least conditionally conjugate priors may not be applicable to the pseudo likelihood necessitating the development of entirely new sampling schemes.

3. Posterior inference based on pseudo representative samples

We now propose an alternative framework for carrying out posterior inference when sample weights must be taken into account. We refer to this as Bayesian Weighted Estimation (BWE). It can be understood as a data augmentation approach where the target posterior includes both parameters and pseudo representative samples (hereafter PRS), denoted $\boldsymbol{z} = (z_1, z_2, \ldots, z_n)'$. First, we define a mechanism for simulating $\boldsymbol{z}$ conditional on both the data and weights. This mechanism is denoted $p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{w})$. Simulation based posterior inference is then carried out as if the PRS were the data, i.e. it is based on the posterior $p(\boldsymbol{\theta}|\boldsymbol{z}) \propto p(\boldsymbol{z}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, where $p(\boldsymbol{z}|\boldsymbol{\theta})$ is the likelihood of the parametric model of interest. A natural way to handle randomness in the mechanism for simulating $\boldsymbol{z}$ is to integrate out over $\boldsymbol{z}$. As such, the approach can be summarised by

$$
\begin{aligned}
p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{w}) &= \int_{\boldsymbol{z}} p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{y}, \boldsymbol{w}) d\boldsymbol{z} \\
&= \int_{\boldsymbol{z}} p(\boldsymbol{\theta}|\boldsymbol{z}, \boldsymbol{y}, \boldsymbol{w}) p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{w}) d\boldsymbol{z} \\
&= \int_{\boldsymbol{z}} p(\boldsymbol{\theta}|\boldsymbol{z}) p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{w}) d\boldsymbol{z} \,.
\end{aligned}
$$

The implicit assumption here is that $\boldsymbol{y}$ and $\boldsymbol{w}$ provide no further information about $\boldsymbol{\theta}$ that is not already captured by $\boldsymbol{z}$. Since this integral cannot be evaluated analytically, the objective is to obtain a Monte Carlo sample of $(\boldsymbol{\theta}^\top, \boldsymbol{z}^\top)^\top$ from $p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{y}, \boldsymbol{w})$.

Ultimately, inference will depend on two choices. The first is the mechanism for generating a PRS. The second is the method used to draw from the posterior of the parameters given $\boldsymbol{z}$, which will depend on the parametric model in question. We now discuss each of these in turn.

3.1. Generating a pseudo representative sample

One way to generate a PRS is to draw a sample of size n from the (weighted) empirical distribution of the data. In our context, that is a discrete distribution with domain $\{y_1, y_2, \ldots, y_n\}$ and with probabilities $\tilde{w}_1, \tilde{w}_2, \ldots, \tilde{w}_n$, where $\tilde{w}_i$ is the normalized weight $\tilde{w}_i = w_i/n$. In the event that all weights are equal this is identical to sampling with replacement, a scheme commonly used in bootstrapping. However, this mechanism for generating the PRS potentially suffers from a number of shortcomings. First, the empirical distribution function is merely an estimate for the process generating a representative sample and uncertainty around this estimate is not explicitly taken into account. Second, simply drawing from the empirical distribution function does not correct for other issues such as a finite population size. The extent to which these factors are a major issue in practice will be investigated in a simulated setting.

To overcome these issues we consider alternatives motivated by the literature on the Bayesian bootstrap (Rubin 1981) and more specifically its weighted version (Lo 1993). In this literature, a distribution is placed on all possible distributions. The empirical distribution function is merely a single realisation from this meta-distribution and equivalent to the posterior mode. Simulation algorithms for the Bayesian Bootstrap rely on Polya's urn schemes which in our context provide a framework for generating a PRS. Specifically we will adopt the algorithm discussed in Dong, Elliott & Raghunathan (2014a) that builds on earlier work by Cohen (1997). This is tailored to the case where survey weights are available and where population size N is finite. This algorithm, which we will refer to as the Weighted Finite Population Bayesian Bootstrap (WFPBB), is summarised below as Algorithm 1.

Dong, Elliott & Raghunathan (2014b) provides extensions to this algorithm that deal with a wide variety of sampling methodologies including cluster-based and stratified sampling. However, to the best of our knowledge these methods have only been applied to find the sampling distribution of a simple statistic of the data. We now discuss how these algorithms can be integrated, in a modular fashion, with simulation based Bayesian inference for a potentially complicated parametric model.

3.2. Simulation based inference

Once an algorithm is chosen for simulating $\boldsymbol{z}$ all that remains is to conduct inference as if pseudo representative samples were actual data. In some cases it is possible to directly draw from $p(\boldsymbol{\theta}|\boldsymbol{z})$ in which case Algorithm 2, described below, can be used.

Since all draws are independent, these steps can be carried out in a sequential or parallel fashion. The class of models for which direct draws from the posterior are possible is limited. However, we consider one such case in Simulation 1 of the following section. In the more

Algorithm 1 Weighted Finite Population Bayesian Bootstrap Dong, Elliott & Raghunathan (2014a)

.

```
procedure WFPBB(y, w̃, N, n).
    l_i ← 0 ∀i = 1, ... n;
    for k = 1 : N − n do
        Letting N* = (N − n)/n, draw y*_k such that y*_k = y_i with probability
```

$$\frac{\tilde{w}_i - 1 + l_i N^*}{N - n + (k-1) \times N^*},$$

```
        if y*_k = y_i then
            l_i ← l_i + 1;
        end if
    end for
    Stack (y_1, y_2, ..., y_n) and (y*_1, y*_2, ..., y*_{N−n}) to form a pseudo population;
    Randomly, draw a sample of size n from the pseudo population;
end procedure
```

Algorithm 2 Direct Posterior Draws with PRS.

```
procedure DPD-PRS(y, w̃, M).
    for i = 1 : M do                                  ▷ This loop can be done in parallel
        Draw z^[i] from p(z|y, w);
        Draw θ^[i] from p(θ|z^[i]);
    end for
end procedure
```

likely event where posterior inference is only possible via MCMC we consider two possible solutions.

3.2.1. Sequential algorithm

Consider that the aim is to construct a Markov chain that converges to a target density $p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{y}, \boldsymbol{w})$. Note that the $(\boldsymbol{y}^\top, \boldsymbol{w}^\top)^\top$, are conditioned on throughout. However, this is suppressed for ease of notation. One option is a Metropolis within Gibbs scheme that draws from $p(\boldsymbol{\theta}|\boldsymbol{z})$ and $p(\boldsymbol{z}|\boldsymbol{\theta})$. The exact method for drawing $p(\boldsymbol{\theta}|\boldsymbol{z})$ will be context specific but can be built up in the usual modular fashion of MCMC. For instance, $\boldsymbol{\theta}$ can be partitioned into blocks some of which are themselves sampled using a Metropolis Hastings step. Of more interest is the proposal for $p(\boldsymbol{z}|\boldsymbol{\theta})$, for which one option is any mechanism for drawing a PRS, as described in Section 3. Letting $\boldsymbol{z}^* \sim p(\boldsymbol{z})$ be the proposed value and $(\boldsymbol{z}^\top, \boldsymbol{\theta}^\top)^\top$, be the current state of the Markov chain, the usual acceptance probability in the Metropolis Hastings algorithm is given by

$$\alpha = \min\left(1, \frac{p(\boldsymbol{z}^*|\boldsymbol{\theta})p(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{\theta})p(\boldsymbol{z}^*)}\right).$$

For some PRS generating mechanisms, such as the empirical distribution, the density $p(\boldsymbol{z})$ is easy to compute. For more complicated mechanisms, such as the finite population Bayesian bootstrap, it is not so straightforward. In this case, it is instructive to manipulate the acceptance ratio as follows:

$$\begin{aligned}\frac{p(\boldsymbol{z}^*|\boldsymbol{\theta})p(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{\theta})p(\boldsymbol{z}^*)} &= \frac{p(\boldsymbol{z}^*,\boldsymbol{\theta})p(\boldsymbol{\theta})p(\boldsymbol{z})}{p(\boldsymbol{z},\boldsymbol{\theta})p(\boldsymbol{\theta})p(\boldsymbol{z}^*)}\\ &= \frac{p(\boldsymbol{\theta}|\boldsymbol{z}^*)}{p(\boldsymbol{\theta}|\boldsymbol{z})}.\end{aligned}$$

This is equivalent to the ratio of posteriors. Note that although in Bayesian inference the normalising constant of the posterior can usually be ignored, that does not apply here since the pseudo representative sample (i.e. the data) is different on the numerator and denominator.

Since both the sequential algorithm and the approach using direct posterior draws are limited in their application we propose an alternative that can be used with any mechanism for generating a PRS and that exploits the potential of parallel computing.

3.2.2. Parallel algorithm

The most flexible algorithm that we propose is one that is well suited to modern parallel computing environments. This involves simulating J pseudo representative samples $\boldsymbol{z}^{[1]},\ldots,\boldsymbol{z}^{[J]}$. For each PRS we can independently simulate an MCMC chain, obtaining M iterations of $\boldsymbol{\theta}$ after a burn-in is discarded and the chain is thinned. This yields a total of $J \times M$ iterates of $\boldsymbol{\theta}$. This procedure is summarised as Algorithm 3 below.

Algorithm 3 Parallel MCMC with PRS.

```
1: procedure MCMC-PRS(y, w̃, M, J).
2:     for i = 1 : J do                      ▷ This loop can be done in parallel
3:         Draw z^[j] from p(z|y, w);
4:         for i = 1 : M do                  ▷ This loop must be done sequentially
5:             Draw θ^[i] from p(θ|z^[j]);
6:         end for
7:     end for
8: end procedure
```

The usual posterior inference can be carried out on this sample of $\boldsymbol{\theta}$. For instance all posterior expectations can be approximated by sample means while credible intervals can be obtained by looking at quantiles of the iterates of $\boldsymbol{\theta}$. The choice of J and M can be tuned depending on the number of cores available in a parallel computing environment and on the mixing performance of the chain. The performance of this approach will be thoroughly investigated in the second part of the following simulation study.

4. Simulation study

In this section we describe two simulation studies that serve dual purposes – to illustrate how the Bayesian weighted estimator is implemented in two specific cases, and to compare the sampling-theory performance of a variety of weighted and unweighted Bayesian and sampling theory estimators. In the first experiment the response variable Y is assumed to follow a normal distribution, while in the second experiment Y is assumed to follow a mixture of gamma distributions. To obtain weights we introduce a normally distributed selection variable X, where dependence between X and Y is induced via a Gaussian copula. The probability that a value of the response variable is observed depends on the selection variable via a probit link function. In both cases we assume that the weights derived from probabilities computed using the probit function are observed, but realisations of X that are used to compute the probabilities and weights are not observed.

4.1. Simulation 1: normal response

When both Y and X are marginally Gaussian and bound by a Gaussian copula the values have a bivariate normal distribution

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim \mathrm{BVN}\left(\begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_x^2 \end{pmatrix}\right).$$

The variable Y is a response variable; we are interested in estimating its mean μ_y and variance σ_y^2. The variable X is a selection variable. When a sample is taken from the population, the X-value for a member of the population determines the probability of selecting that member of the population into the sample. Specifically, we assume that Y_s is selected into the sample if and only if $I_s = 1$, where

$$\Pr\left(I_s = 1 | Y_s, X_s\right) = \Pr\left(I_s = 1 | X_s\right) = \pi_s = \Phi\left(\beta_0 + \beta_1 X_s\right)\,,$$

with $\Phi\left(\cdot\right)$ denoting the cumulative distribution function of a standard normal distribution. When a member of the population is selected into the sample, we observe Y_s and a weight w_s assumed to be such that $w_s \propto 1/\pi_s$, but we do not observe X_s. The selected sample is denoted as $\left(\boldsymbol{y}^\top, \boldsymbol{w}^\top\right)^\top$. Scaling the weights so that they sum to the sample size, we have

$w_s = n\pi_s^{-1}/\sum_{t=1}^{n}\pi_t^{-1}$. The normalized sampling weights are given by $\widetilde{w}_s = w_s/\sum_{t=1}^{n} w_t$. The objective is to use $\left(\boldsymbol{y}^\top, \widetilde{\boldsymbol{w}^\top}\right)^\top$ to estimate μ_y and σ_y^2.

The simulation setup we used is as follows: $N = 100,000$ values of (Y_s, X_s) are generated as a finite population, with $\mu_x = 0$, $\sigma_x^2 = 9$, $\mu_y = 10$, and $\sigma_y^2 = \{4, 16, 100\}$. A sample drawn from this population will be representative, in the sense that each population value of Y has an equal chance of being selected, if $\rho = 0$ or $\beta_1 = 0$. Thus, for $\beta_1 \neq 0$, the value of ρ controls the representativeness of the sample. Three different variances are used because the impact of an unrepresentative sample is potentially worse for larger variances. With larger variances, extreme values of Y will be systematically omitted from the sample. To obtain an observed sample, each population pair (Y_s, X_s) is assigned a probability π_s from the probit function and selected with probability π_s. The probit function parameters used for this exercise were $\beta_0 = \{-1.8, -2.7\}$ and $\beta_1 = 0.1$. For a given β_1, the setting for β_0 controls the sample size; $\beta_0 = -1.8$ leads to a sample of approximately 4000, and, for $\beta_0 = -2.7$, $n \approx 500$.

In Figure 1 we plot histograms for examples of samples of Y generated with $\beta_0 = -2.7$, $\beta_1 = 0.1$, $\sigma_y^2 = 16$ and the three values $\rho = \{0, 0.2, 0.8\}$. When $\rho = 0$, the sample is "representative" and the histogram is centred close to the true value $\mu_y = 10$. Increasing ρ to 0.2 moves the distribution slightly to the right centering it at $\overline{y} = 10.65$. A further increase in ρ to 0.8 leads to a substantial shift, centering the distribution at $\overline{y} = 12.71$.

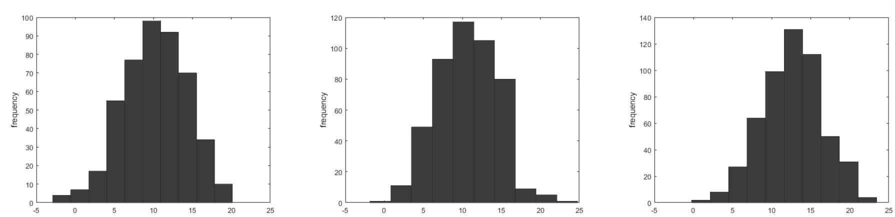

Figure 1. Histograms of selected samples of Y. For the left panel (no selection) $\rho = 0$, $\overline{y} = 10.22$, $s_y = 4.05$, for the middle panel $\rho = 0.2$, $\overline{y} = 10.65$, $s_y = 3.80$, $n = 471$ and for the right panel $\rho = 0.8$, $\overline{y} = 12.71$, $s_y = 3.88$, $n = 528$.

We use 250 Monte Carlo replications to examine the performance of four Bayesian and one sampling theory estimators for μ_y. For each estimator results are reported for:

1. The average of estimates for μ_y;
2. The average of the variance estimates for each estimator for μ_y – either the relevant sampling theory estimator or the posterior variance for μ_y;
3. The coverage of 95% interval estimates for μ_y constructed using the estimates from (2) and (3).

Details of the estimators follow. Derivations are provided in the online supplementary material.

1. **Pseudo MLE (PMLE):** The closed form solutions are $\widehat{\mu}_{y,PMLE} = (1/n)\sum_{s=1}^{n} w_s y_s$ and $\widehat{\sigma}^2_{y,PMLE} = \left(1/n^2\right)\sum_{s=1}^{n} w_s^2 \left(y_s - \widehat{\mu}_{y,PMLE}\right)^2$.
2. **Unweighted Bayesian (UBE):** Using the non-informative joint prior distribution $p\left(\mu_y, \sigma_y^2\right) = 1/\sigma_y^2$, we obtain the marginal posteriors $\sigma_y^2|\boldsymbol{y} \sim \text{IG}\left(v/2, v\widetilde{s}^2/2\right)$, and $\mu_y|\boldsymbol{y} \sim \text{t}\left(\overline{y}, v\widetilde{s}^2/(v-2)n\right)$, where $v = n-1$ and $\widetilde{s}^2 = v^{-1}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2$. The posterior mean $\overline{y}$ is used as a point estimate for μ_y, and the posterior variance for μ_y is used as the variance estimate for $\overline{y}$. Except for a degrees of freedom correction which is inconsequential for the sample sizes considered here, the posterior mean and variance are identical to the mean and variance for an unweighted MLE. Thus, the results for the UBE are also indicative of those for unweighted MLE.
3. **Bayesian Pseudo Posterior (BPPE):** From the joint pseudo posterior density $\mu_y, \sigma_y^2|\boldsymbol{y}, \boldsymbol{w} \sim \sigma_y^{-2}\prod_{s=1}^{n}\left(\phi\left(y_s; \mu_y, \sigma_y^2\right)\right)^{w_s}$, where $\phi(y; a, b)$ is the normal density with mean a and variance b, we obtain the marginal distributions $\widetilde{p}\left(\sigma_y^2|\boldsymbol{y}, \boldsymbol{w}\right)$ which is distributed $\text{IG}\left(v/2, v\widetilde{s}^{*2}/2\right)$ and $\widetilde{p}\left(\mu_y|\boldsymbol{y}, \boldsymbol{w}\right)$, which is distributed $\text{t}\left(\overline{y}^*, v\widetilde{s}^{*2}/(v-2)n\right)$, where $\overline{y}^* = n^{-1}\sum_{s=1}^{n} w_s y_s$ and $\widetilde{s}^{*2} = v^{-1}\sum_{s=1}^{n} w_s\left(y_s - \overline{y}^*\right)^2$. The posterior mean $\overline{y}^*$ is used as a point estimate for μ_y, and the posterior variance of μ_y is used as the variance of this estimate.
4. **Bayesian Weighted (BWE):** Adapting Algorithm 3 in Section 3.2, the first step is to draw PRS $\boldsymbol{z}$ from $p\left(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{w}\right)$ and is discussed in detail in Section 3.1. We now discuss the second step, drawing $\boldsymbol{\theta}^{(i)}$ conditional on the PRS $\boldsymbol{z}^{(i)}$ from $p\left(\boldsymbol{\theta}|\boldsymbol{z}^{(i)}\right)$ at the iteration i. First, we compute $\overline{z}^{(i)} = n^{-1}\sum_{s=1}^{n} z_s^{(i)}$ and $\widetilde{s}^{2(i)} = (n-1)^{-1}\sum_{s=1}^{n}\left(z_s^{(i)} - \overline{z}^{(i)}\right)^2$. Then, we draw $\sigma_y^{2(i)}$ from $\text{IG}\left(v/2, v\widetilde{s}^{2(i)}/2\right)$, where $v = n-1$ and $\mu_y^{(i)}$ from $p\left(\mu_y^{(i)}|\sigma_y^{2(i)}, \boldsymbol{z}^{(i)}\right)$.
   We use BWE-EDF to refer to the algorithm that generates a PRS by drawing a sample of size n from the (weighted) empirical distribution of the data and BWE-WFPBB for the algorithm that generates a PRS by using Weighted Finite Population Bayesian Bootstrap in Algorithm 1. A total of $M = 2000$ posterior draws were generated. The posterior draws were used to estimate posterior means and variances for $\left(\mu_y, \sigma_y^2\right)$. The $\left(\mu_y, \sigma_y^2\right)^{(1)}, \ldots, \left(\mu_y, \sigma_y^2\right)^{(M)}$ approximate draws from the posterior distribution $p\left(\mu_y, \sigma_y^2|\boldsymbol{y}, \widetilde{\boldsymbol{w}}\right)$. For estimates of the posterior mean and variance of μ_y, we can use $\widehat{\mu}_y = M^{-1}\sum_{i=1}^{M}\overline{z}^{(i)}$ and $\widehat{\sigma}^2_\mu = M^{-1}\sum_{i=1}^{M}\sigma_y^{2(i)}/n + M^{-1}\sum_{i=1}^{M}\left(\overline{z}^{(i)} - \widehat{\mu}_y\right)^2$.

The means of the point estimates for $\widehat{\mu}_y$ and its variance $\widehat{\sigma}^2_\mu$ were calculated over $R = 250$ replications for each method. We use $\overline{\mu}_y = (1/R)\sum_{r=1}^{R} w_r\widehat{\mu}_{y,r}$ to denote the average of the estimates of μ_y and $\overline{\sigma}^2_\mu = (1/R)\sum_{r=1}^{R} w_r\widehat{\sigma}^2_{\mu,r}$ to denote the average of the estimates of the variance of $\widehat{\mu}_y$ where $\widehat{\sigma}^2_{\mu,r}$ is the posterior variance of μ_y under Bayesian frameworks and the variance of $\widehat{\mu}_y$ under a frequentist framework.

In Tables 1 to 5, we report the results for $\overline{\mu}_y$ and $\overline{\sigma}^2_\mu$ from the various estimators, together with the coverage of 95% Bayesian credible intervals and 95% frequentist confidence intervals. A coverage less than 95% suggests that the variance of an estimate for μ_y is biased downwards and a coverage greater than 95% suggests the variance estimate is biased upwards. Table 1 contains results for the case where Y and X are uncorrelated $(\rho = 0)$. Tables 2 and 3 contain results for a large observed sample size, high and low correlation $(\rho = 0.8, 0.2)$ and different values for the variance of Y $\left(\sigma^2_y = 4, 16, 100\right)$. Tables 4 and 5 contain the corresponding results for a small observed sample size. We observe that:

1. The estimates for μ_y from PMLE, BPPE, and both BWE-EDF, and BWE-WFPBB, the estimators which utilize the weights, are close to the true value $\mu_y = 10$, even when the observed sample size is only approximately 500, suggesting that any bias in these estimators is negligible. The unweighted estimator is biased, however. The amount of bias depends on three things: the true variance of Y, the degree of correlation between Y and X, and the sample size. The higher the degree of correlation ρ, the larger the true variance of Y, or the smaller the observed sample size, the larger the bias of the unweighted estimator.
2. From Table 1 where $\rho = 0$, the mean of the unweighted estimates for the parameter μ_y is close to the true value suggesting that when Y is not correlated with X, the unweighted estimator is unbiased. The PMLE, BWE-EDF, and BWE-WFPBB have higher variance estimates on average compared to UBE, reflecting the effect of unnecessary complexity.
3. The average of the variance estimates over the replications, $\overline{\sigma}^2_\mu$, is always smaller for BPPE compared to PMLE and BWE (Tables 2 to 5). These smaller variance estimates for BPPE lead to interval estimate coverage that is smaller than the PMLE, BWE-EDF, and BWE-WFPBB. Using BPPE, the variance of the estimates is underestimated since the wrong variance matrix is employed. PMLE uses the robust "sandwich estimator" to correctly estimate the variance matrix. Both BWE-EDF and BWE-WFPBB estimators integrate out across pseudo representative samples $\boldsymbol{z}$ to their posterior distributions.
4. Increasing the variance σ^2_y increases the average variance $\overline{\sigma}^2_\mu$, but it does not change coverage.
5. In most cases, the coverage of BWE-WFPBB is comparable in magnitude to the 95% confidence intervals of PMLE. The coverage of BWE-EDF is slightly lower than the BWE-WFPBB. The averages of the variances of the estimates are also quite comparable for PMLE and both BWE estimators. Thus, the BWE's posterior variance

Table 1. Estimates for parameter μ_y with true values $\mu_y = 10$, $\rho = 0$, and $n \approx 4000$.

Case		UBE	PMLE	BPPE	BWE-EDF	BWE-WFPBB
	$\overline{\mu}_y$	10.0222	10.0143	10.0143	10.0139	10.0147
$\sigma_y^2 = 100$	$\overline{\sigma}_\mu^2$	0.0235	0.0364	0.0235	0.0470	0.0704
	coverage	0.9400	0.9560	0.9000	0.9600	0.9800

Table 2. Estimates for parameter μ_y with true values $\mu_y = 10$, $\rho = 0.8$, and $n \approx 4000$.

Case		UBE	PMLE	BPPE	BWE-EDF	BWE-WFPBB
	$\overline{\mu}_y$	14.8991	10.0153	10.0153	10.0140	10.0150
$\sigma_y^2 = 100$	$\overline{\sigma}_\mu^2$	0.0225	0.0494	0.0235	0.0465	0.0693
	coverage	0.0000	0.9440	0.8160	0.9440	0.9720
	$\overline{\mu}_y$	11.9595	10.0059	10.0059	10.0054	10.0059
$\sigma_y^2 = 16$	$\overline{\sigma}_\mu^2$	0.0036	0.0079	0.0038	0.0075	0.0111
	coverage	0.0000	0.9440	0.8160	0.9440	0.9720
	$\overline{\mu}_y$	10.9795	10.0027	10.0027	10.0024	10.0026
$\sigma_y^2 = 4$	$\overline{\sigma}_\mu^2$	0.0008	0.0020	0.0009	0.0019	0.0028
	coverage	0.0000	0.9440	0.8200	0.9440	0.9720

Table 3. Estimates for parameter μ_y with true values $\mu_y = 10$, $\rho = 0.2$, and $n \approx 4000$.

Case		UBE	PMLE	BPPE	BWE-EDF	BWE-WFPBB
	$\overline{\mu}_y$	11.2429	10.0159	10.0159	10.0168	10.0171
$\sigma_y^2 = 100$	$\overline{\sigma}_\mu^2$	0.0235	0.0371	0.0235	0.0466	0.0700
	coverage	0.0000	0.9440	0.8840	0.9640	0.9880
	$\overline{\mu}_y$	10.4977	10.0068	10.0068	10.0072	10.0073
$\sigma_y^2 = 16$	$\overline{\sigma}_\mu^2$	0.0038	0.0059	0.0038	0.0075	0.0112
	coverage	0.0000	0.9440	0.8840	0.9640	0.9880
	$\overline{\mu}_y$	10.2487	10.0030	10.0030	10.0032	10.0033
$\sigma_y^2 = 4$	$\overline{\sigma}_\mu^2$	0.0009	0.0015	0.0009	0.0019	0.0028
	coverage	0.0000	0.9440	0.8840	0.9640	0.9880

can be thought of as a Bayesian way of correcting the posterior variance when sampling weights are taken into account.

Table 4. Estimates for parameter μ_y with true values $\mu_y = 10$, $\rho = 0.8$, and $n \approx 500$.

Case		UBE	PMLE	BPPE	BWE-EDF	BWE-WFPBB
	$\overline{\mu}_y$	16.6339	9.9127	9.9127	9.9115	9.9146
$\sigma_y^2 = 100$	$\overline{\sigma}_\mu^2$	0.1982	0.8751	0.2096	0.4204	0.6180
	coverage	0.0000	0.9440	0.6880	0.8400	0.8960
	$\overline{\mu}_y$	12.6533	9.9740	9.9740	9.9735	9.9748
$\sigma_y^2 = 16$	$\overline{\sigma}_\mu^2$	0.0317	0.1380	0.0333	0.0668	0.0983
	coverage	0.0000	0.9440	0.7000	0.8480	0.9000
	$\overline{\mu}_y$	11.3269	9.9822	9.9822	9.9820	9.9825
$\sigma_y^2 = 4$	$\overline{\sigma}_\mu^2$	0.0079	0.0350	0.0084	0.0168	0.0247
	coverage	0.0000	0.9440	0.6920	0.8400	0.8960

Table 5. Estimates for parameter μ_y with true values $\mu_y = 10$, $\rho = 0.2$, and $n \approx 500$.

Case		UBE	PMLE	BPPE	BWE-EDF	BWE-WFPBB
	$\overline{\mu}_y$	11.6386	9.9927	9.9927	9.9938	9.9966
$\sigma_y^2 = 100$	$\overline{\sigma}_\mu^2$	0.2058	0.4932	0.2073	0.4144	0.6220
	coverage	0.0480	0.9320	0.7720	0.8920	0.9560
	$\overline{\mu}_y$	10.6557	10.0008	10.0008	10.0012	10.0024
$\sigma_y^2 = 16$	$\overline{\sigma}_\mu^2$	0.0329	0.0791	0.0332	0.0663	0.0996
	coverage	0.0480	0.9360	0.7760	0.8960	0.9600
	$\overline{\mu}_y$	10.3265	9.9976	9.9976	9.9977	9.9983
$\sigma_y^2 = 4$	$\overline{\sigma}_\mu^2$	0.0082	0.0197	0.0083	0.0166	0.0249
	coverage	0.0480	0.9320	0.7720	0.8960	0.9600

4.2. Simulation 2: finite gamma mixture

In this section we illustrate how to integrate the Bayesian weighted estimator within an MCMC algorithm for estimation of the parameters of a more complex model. We consider a finite mixture of gamma densities with two components. The procedure can be readily extended to the case of K components. We assume that the population distribution for a response variable Y can be described by the density

$$p\left(y|\xi,\boldsymbol{\mu},\boldsymbol{v}\right) = \xi \mathrm{G}\left(y|v_{1,}\mu_1\right) + \left(1-\xi\right)\mathrm{G}\left(y|v_{2,}\mu_2\right),$$

where v_k is the shape parameter and μ_k is the mean of the gamma density

$$\mathrm{G}\left(y|v_k,\mu_k\right) = \frac{\left(v_k/\mu_k\right)^{v_k}}{\Gamma\left(v_k\right)} y^{v_k-1}\exp\left(-\frac{v_k}{\mu_k}y\right).$$

The marginal distribution of the selection variable X is assumed to be N $\left(\mu_X, \sigma_X^2\right)$ as in the first simulation and a bivariate Gaussian copula is employed to construct a joint distribution between X and Y. Steps to generate a population for (Y, X) are given in Section D of the supplementary material. A similar set up to simulation 1 is used to select the sample and to compute the sampling weights. For the estimation of $(\xi, \mu_1, \mu_2, v_1, v_2)^\top$, we assume that only the sampling weights and the sample observations $\boldsymbol{y}$ are observed.

The true parameters for the mixture of gamma densities were set as follows: $\xi = 0.6$, $\mu_1 = 208$, $\mu_2 = 700$, $v_1 = 3$ and $v_2 = 2$. Those for X were $\mu_X = 0$ and $\sigma_X^2 = 9$. The correlation ρ was set to be $\{0, 0.2, 0.5, 0.8\}$. The probit function parameters used for this exercise were $\beta_0 = \{-1.2, -1.8\}$ and $\beta_1 = 0.1$. For a given β_1, the setting for β_0 controls the sample size; $\beta_0 = -1.2$ leads to a sample of approximately 12% of the whole finite population distribution and $\beta_0 = -1.8$ leads to a sample of approximately 4% of the whole finite population distribution. The total number of Monte Carlo replications R was set at 250.

The MCMC algorithm used to estimate the model combines that suggested by Wiper, Insua & Ruggeri (2001), with our proposal for including the weights. We describe it in terms of a general model with K components. The priors employed by Wiper, Insua & Ruggeri (2001) are a Dirichlet prior for $\boldsymbol{\xi}$

$$p\left(\boldsymbol{\xi}\right) \propto \xi_1^{\varphi_1 - 1} \xi_2^{\varphi_2 - 1} ... \xi_K^{\varphi_K - 1},$$

an inverted gamma prior IG (α_k, β_k) for μ_k with density,

$$p\left(\mu_k\right) \propto \left(\mu_k\right)^{-(\alpha_k + 1)} \exp\left(-\frac{\beta_k}{\mu_k}\right),$$

and an exponential prior for v_k

$$p\left(v_k\right) \propto \exp\left(-\lambda v_k\right).$$

Adapting Algorithm 3 in Section 3.2.2, the step to draw PRS $\boldsymbol{z}^{(j)}$ from $p\left(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{w}\right)$ is discussed in detail in Section 3.1. We now discuss the second step, drawing $\boldsymbol{\theta}^{(i)} = \left(\boldsymbol{\xi}^{(i)}, \boldsymbol{v}^{(i)}, \boldsymbol{\mu}^{(i)}\right)$ conditional on the PRS $\boldsymbol{z}^{(j)}$ from $p\left(\boldsymbol{\xi}^{(i)}, \boldsymbol{v}^{(i)}, \boldsymbol{\mu}^{(i)}|\boldsymbol{z}^{(j)}\right)$ at the iteration i, where $\boldsymbol{\xi}^{(i)} = \left(\xi_1^{(i)}, ..., \xi_K^{(i)}\right)^\top$, $\boldsymbol{\mu}^{(i)} = \left(\mu_1^{(i)}, ..., \mu_K^{(i)}\right)^\top$, and $\boldsymbol{v}^{(i)} = \left(v_1^{(i)}, ..., v_K^{(i)}\right)^\top$. The steps of drawing $\boldsymbol{\theta}^{(i)}$ for $i = 1, ..., M$ are summarised as

1. Generate $\left(\boldsymbol{d}_s^{(i)}|\boldsymbol{\xi}^{(i)}, \boldsymbol{v}^{(i)}, \boldsymbol{\mu}^{(i)}, \boldsymbol{z}^{(j)}\right)$ for $s = 1, ..., n$, where $\boldsymbol{d}_s = (d_{s1}, ..., d_{sK})$, and d_{sk} is an indicator variable equal to 1 if the sth observation is identified as coming

from the kth component of the mixture according to the probability

$$p\left(d_{sk}=1|\boldsymbol{z},\boldsymbol{\xi},\boldsymbol{\mu},\boldsymbol{v}\right)=\frac{p_{sk}}{p_{s1}+...+p_{sK}},$$

where

$$p_{sk}=\xi_k\frac{\left(v_k/\mu_k\right)^{v_k}}{\Gamma\left(v_k\right)}z_s^{v_k-1}\exp\left(-\frac{v_k}{\mu_k}z_s\right).$$

Let $\boldsymbol{D}$ be the $(n\times K)$ matrix of components d_{sk} and $n_k=\sum_{s=1}^{n}w_s d_{sk}$.

2. Generate $\left(\boldsymbol{\xi}^{(i)}|\boldsymbol{D}^{(i)},\boldsymbol{\mu}^{(i-1)},\boldsymbol{v}^{(i-1)},\boldsymbol{z}^{(j)}\right)$ from the Dirichlet distribution

$$\boldsymbol{\xi}|\boldsymbol{z},\boldsymbol{D},\boldsymbol{\mu},\boldsymbol{v}\sim \mathrm{D}\left(\boldsymbol{\varphi}+\boldsymbol{n}\right),$$

where $\boldsymbol{n}^{\top}=(n_1,...,n_K)$ and $\boldsymbol{\varphi}^{\top}=(\varphi_1,...,\varphi_K)$.

3. Generate $\left(\mu_k^{(i)}|\boldsymbol{D}^{(i)},\boldsymbol{\xi}^{(i)},\boldsymbol{v}^{(i-1)},\boldsymbol{z}^{(j)}\right)$ for $k=1,...,K$ from the inverted gamma density

$$\mu_k|\boldsymbol{z},\boldsymbol{D},\boldsymbol{v},\boldsymbol{\xi}\sim \mathrm{IG}\left(\alpha_k+n_k v_k,\beta_k+S_k v_k\right),$$

where $S_k=\boldsymbol{w}_{s=1}^{n}d_{sk}z_s$.

4. Generate $\left(v_k^{(i)}|\boldsymbol{D}^{(i)},\boldsymbol{\xi}^{(i)},\boldsymbol{\mu}^{(i)},\boldsymbol{z}^{(j)}\right)$, for $k=1,...,K$ from

$$p\left(v_k|\boldsymbol{z},\boldsymbol{D},\boldsymbol{\mu},\boldsymbol{\xi}\right)\propto\frac{v_k^{n_k v_k}}{\left[\Gamma\left(v_k\right)\right]^{n_k}}\exp\left\{-v_k\left(\lambda+\frac{S_k}{\mu_k}+n_k\log\mu_k-P_k\right)\right\},$$

where $P_k=\sum_{s=1}^{n}w_s d_{sk}\log z_s$. Values are drawn from this density using a Metropolis step with a gamma candidate generating function $v_k^{*(i)}\sim \mathrm{G}\left(r_k,r_k/v_k^{(i-1)}\right)$ with r_k chosen by experimentation to obtain a reasonable acceptance rate.

5. For identification, order the elements according to $\mu_1<...<\mu_K$.

We use the abbreviations BWE-EDF and BWE-WFPBB in the same manner as Section 4.1 but where Algorithm 3 is used. We simulate $J=200$ pseudo representative samples (PRS) $\boldsymbol{z}^{(1)},...,\boldsymbol{z}^{(J)}$. For each PRS we independently simulate an MCMC chain, obtaining a total of $M=5500$ observations on $\boldsymbol{\theta}=\left(\boldsymbol{\mu}^{\top},\boldsymbol{v}^{\top},\xi^{\top}\right)^{\top}$, with the first 500 draws discarded as a burn in, a total of 200×5000 iterates of $\boldsymbol{\theta}$ for each replication.

A total of $R=250$ Monte Carlo replications were taken, and for a sample of the replications, the observations were plotted to confirm the convergence of the Markov chains. Following Wiper, Insua & Ruggeri (2001), relatively noninformative priors were used with the parameter settings $\varphi_1=\varphi_2=1$, $\alpha_1=\alpha_2=2.2$, $\beta_1=40$, $\beta_2=80$, and $\lambda_1=\lambda_2=0.01$. We also impose a priori restriction $\mu_1<\mu_2$ for identification of the mixture components. If the objective is estimation of the overall mixture distribution and not the individual parameters, as is the case for our empirical example in the next section, then the identification restriction is unnecessary (Geweke 2007).

In Tables 6 to 9, we report the averages of the posterior means $\overline{\theta}$, coverage of the 95% Bayesian credible intervals, and the averages of the posterior variances $\overline{\sigma}^2_\theta$ from the various estimators. Table 6 contains results for the case where Y and X are uncorrelated $(\rho = 0)$. Tables 7 to 9 contain results for small and large observed sample sizes, with correlations $\rho = 0.2$, 0.5, and 0.8.

We observe the following:

1. From Table 6 where the correlation $\rho = 0$, the components of $\boldsymbol{\theta}$ are close to their true counterparts for the UBE and both BWE-EDF and BWE-WFPBB. It suggests that when Y is not correlated with X, the unweighted estimator is unbiased. The interval estimate of UBE has coverage that is close to the nominal 95%, but the interval estimates of the BWE-EDF and BWE-WFPBB have coverage that is higher than the nominal 95%. Both BWE-EDF and BWE-WFPBB have higher variance estimates on average compared to UBE, reflecting the effect of unnecessary complexity.
2. From Tables 7 to 9, the components of $\boldsymbol{\theta}$ are close to their true counterparts suggesting that any bias in BWE-EDF and BWE-WFPBB is negligible for both sample sizes. The unweighted estimator is biased, however. The higher the degree of correlation ρ, or the smaller the observed sample size, the larger the bias of the unweighted estimator. As shown in Figure 2, the true density and the estimated densities using the values $\overline{\boldsymbol{\theta}}$ from BWE-EDF and BWE-WFPBB, with $\rho = 0.8$ and $\beta_0 = -1.8$, are indistinguishable, but the estimated density from UBE is clearly far from the true density.
3. With the exception of μ_2, the averages of the posterior variances are relatively small, implying estimation is relatively precise. The BWE-WFPBB estimators have larger averages of posterior variances, $\overline{\sigma}^2_\theta$ for all cases compared to BWE-EDF.
4. Tables 7 to 9 show that the BWE-WFPBB and BWE-EDF coverage of the 95% credible intervals for all parameters $\overline{\boldsymbol{\theta}}$ is quite close to 0.95 when $\rho = 0.8$, but they seem to have over coverage for $\rho = 0$, 0.2, and 0.5.

Thus, we conclude the BWE algorithms work not only for the simple model described in the first simulation, but also for estimating unknown parameters of gamma mixture models. It is a very general algorithm that can be easily extended to integrate with the usual MCMC algorithms, such as the Metropolis-Hastings, Gibbs sampling, and Metropolis-within-Gibbs sampling schemes.

Table 6. Simulation 2: Finite gamma mixture with $\rho = 0$, true values $\xi = 0.6$, $\mu_1 = 208$, $\mu_2 = 700$, $v_1 = 3$ and $v_2 = 2$.

			ξ	μ_1	μ_2	v_1	v_2
	True		0.6000	208.0000	700.0000	3.0000	2.0000
$\beta_0 = -1.2$	UBE	$\overline{\theta}$	0.6060	208.9679	710.6751	3.0114	2.0844
		$\overline{\sigma}^2_\theta$	0.0008	12.5197	1105.60	0.0105	0.0425
		coverage	0.9240	0.9160	0.9280	0.9680	0.9360
	BWE-EDF	$\overline{\theta}$	0.6074	209.2593	714.2218	3.0200	2.1152
		$\overline{\sigma}^2_\theta$	0.0015	25.4134	2094.10	0.0206	0.0859
		coverage	0.9880	0.9880	0.9880	0.9960	0.9960
	BWE-WFPBB	$\overline{\theta}$	0.6084	209.4519	716.6443	3.0254	2.1399
		$\overline{\sigma}^2_\theta$	0.0020	36.1222	2860.60	0.0289	0.1219
		coverage	1.0000	0.9960	0.9960	1.0000	1.0000
$\beta_0 = -1.8$	UBE	$\overline{\theta}$	0.6053	209.1985	714.1303	3.0311	2.1306
		$\overline{\sigma}^2_\theta$	0.0019	35.8193	2725.80	0.0305	0.1165
		coverage	0.9560	0.9480	0.9600	0.9400	0.9600
	BWE-EDF	$\overline{\theta}$	0.6071	209.3585	721.0308	3.0680	2.2223
		$\overline{\sigma}^2_\theta$	0.0034	72.8264	5032.30	0.0639	0.2578
		coverage	0.9720	0.9720	0.9720	0.9840	0.9800
	BWE-WFPBB	$\overline{\theta}$	0.6096	209.8237	726.6071	3.0797	2.2846
		$\overline{\sigma}^2_\theta$	0.0046	108.0590	7008.30	0.0950	0.3953
		coverage	0.9960	0.9880	0.9920	0.9960	0.9920

Table 7. Simulation 2: Finite gamma mixture with $\rho = 0.2$, true values $\xi = 0.6$, $\mu_1 = 208$, $\mu_2 = 700$, $v_1 = 3$ and $v_2 = 2$.

			ξ	μ_1	μ_2	v_1	v_2
	True		0.6000	208.0000	700.0000	3.0000	2.0000
$\beta_0 = -1.2$	UBE	$\overline{\theta}$	0.5734	217.0979	740.5767	3.1366	2.1163
		$\overline{\sigma}^2_\theta$	0.0008	14.7498	1071.50	0.0131	0.0401
		coverage	0.8240	0.2480	0.7800	0.7680	0.9600
	BWE-EDF	$\overline{\theta}$	0.6051	208.7960	711.0733	3.0300	2.0946
		$\overline{\sigma}^2_\theta$	0.0015	25.0760	2053.20	0.0211	0.0815
		coverage	0.9880	0.9960	0.9920	0.9880	0.9920
	BWE-WFPBB	$\overline{\theta}$	0.6061	208.9918	713.2825	3.0360	2.1180
		$\overline{\sigma}^2_\theta$	0.0020	35.4075	2785.10	0.0298	0.1155
		coverage	1.0000	1.0000	1.0000	0.9920	1.0000
$\beta_0 = -1.8$	UBE	$\overline{\theta}$	0.5621	219.6995	750.2276	3.2051	2.1667
		$\overline{\sigma}^2_\theta$	0.0020	42.2860	2505.00	0.0403	0.1018
		coverage	0.8960	0.5320	0.8960	0.8400	0.9680
	BWE-EDF	$\overline{\theta}$	0.6039	208.8690	715.4016	3.0921	2.1966
		$\overline{\sigma}^2_\theta$	0.0033	70.7234	4763.30	0.0649	0.2379
		coverage	0.9840	0.9960	0.9840	0.9560	0.9840
	BWE-WFPBB	$\overline{\theta}$	0.6070	209.4192	721.4411	3.1011	2.2591
		$\overline{\sigma}^2_\theta$	0.0045	104.8445	6689.80	0.0963	0.3660
		coverage	0.9960	0.9960	0.9920	0.9920	0.9880

Table 8. Simulation 2: Finite gamma mixture with $\rho = 0.5$, true values $\xi = 0.6$, $\mu_1 = 208$, $\mu_2 = 700$, $v_1 = 3$ and $v_2 = 2$.

			ξ	μ_1	μ_2	v_1	v_2
	True		0.6000	208.0000	700.0000	3.0000	2.0000
$\beta_0 = -1.2$	UBE	$\overline{\theta}$	0.5250	230.0632	787.5132	3.3626	2.2097
		$\overline{\sigma}^2_\theta$	0.0007	16.9847	843.9421	0.0167	0.0320
		coverage	0.2360	0.0000	0.0520	0.0760	0.7880
	BWE-EDF	$\overline{\theta}$	0.6022	208.7241	708.6371	3.0401	2.0910
		$\overline{\sigma}^2_\theta$	0.0014	24.6815	1965.70	0.0217	0.0774
		coverage	0.9800	0.9840	0.9800	0.9720	0.9720
	BWE-WFPBB	$\overline{\theta}$	0.6033	208.9423	710.8749	3.0459	2.1112
		$\overline{\sigma}^2_\theta$	0.0019	34.9652	2669.80	0.0306	0.1083
		coverage	0.9920	0.9960	0.9920	0.9960	0.9920
$\beta_0 = -1.8$	UBE	$\overline{\theta}$	0.4997	237.0224	815.3224	3.5030	2.2833
		$\overline{\sigma}^2_\theta$	0.0018	55.5087	2121.90	0.0598	0.0882
		coverage	0.3600	0.0040	0.0080	0.2480	0.8640
	BWE-EDF	$\overline{\theta}$	0.5988	209.1928	711.7804	3.0968	2.1667
		$\overline{\sigma}^2_\theta$	0.0033	71.0910	4704.40	0.0671	0.2268
		coverage	0.9680	0.9960	0.9680	0.9520	0.9680
	BWE-WFPBB	$\overline{\theta}$	0.6025	209.7359	718.5456	3.1048	2.2317
		$\overline{\sigma}^2_\theta$	0.0045	106.1416	6697.30	0.0996	0.3527
		coverage	0.9800	0.9960	0.9800	0.9840	0.9760

Table 9. Simulation 2: Finite gamma mixture with $\rho = 0.8$, true values $\xi = 0.6$, $\mu_1 = 208$, $\mu_2 = 700$, $v_1 = 3$ and $v_2 = 2$.

			ξ	μ_1	μ_2	v_1	v_2
	True		0.6000	208.0000	700.0000	3.0000	2.0000
$\beta_0 = -1.2$	UBE	$\overline{\theta}$	0.4793	244.4544	836.5023	3.6288	2.3232
		$\overline{\sigma}^2_\theta$	0.0006	21.0618	757.9739	0.0229	0.0302
		coverage	0.0040	0.0000	0.0000	0.0000	0.3880
	BWE-EDF	$\overline{\theta}$	0.5988	208.6350	704.0435	3.0556	2.0627
		$\overline{\sigma}^2_\theta$	0.0015	24.2958	1951.80	0.0235	0.0757
		coverage	0.9600	0.9760	0.9560	0.9520	0.9480
	BWE-WFPBB	$\overline{\theta}$	0.6007	208.8230	707.1057	3.0584	2.0892
		$\overline{\sigma}^2_\theta$	0.0020	34.8552	2721.40	0.0335	0.1081
		coverage	0.9720	0.9960	0.9720	0.9760	0.9760
$\beta_0 = -1.8$	UBE	$\overline{\theta}$	0.4334	253.7392	877.2529	3.9058	2.3765
		$\overline{\sigma}^2_\theta$	0.0016	68.7337	1772.50	0.0957	0.0712
		coverage	0.0120	0.0000	0.0000	0.0080	0.6240
	BWE-EDF	$\overline{\theta}$	0.5930	208.7730	704.0061	3.1535	2.1104
		$\overline{\sigma}^2_\theta$	0.0032	66.2321	4387.00	0.0747	0.1955
		coverage	0.9160	0.9800	0.9200	0.8920	0.9200
	BWE-WFPBB	$\overline{\theta}$	0.5957	209.2470	709.4283	3.1644	2.1619
		$\overline{\sigma}^2_\theta$	0.0043	98.3922	6134.70	0.1118	0.2951
		coverage	0.9480	0.9960	0.9400	0.9160	0.9360

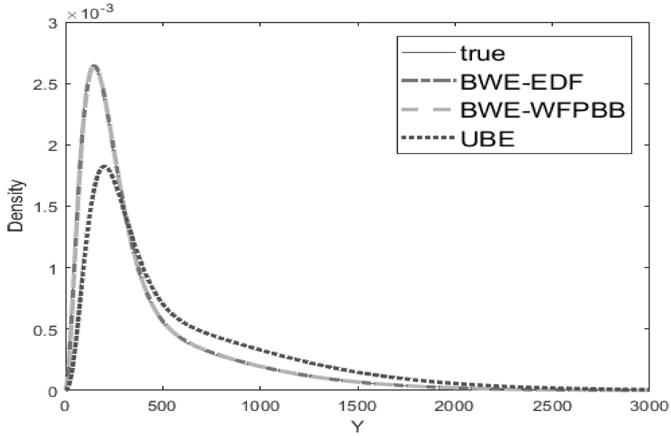

Figure 2. A true gamma mixture density and its estimates from the posterior mean of the BWE-EDF, BWE-WFPBB, and UBE with $\rho = 0.8$, true values $\xi = 0.6$, $\mu_1 = 208$, $\mu_2 = 700$, $v_1 = 3$ and $v_2 = 2$, and $\beta_0 = -1.8$.

5. Application to Australian income distribution

In this section we illustrate our methodology by fitting a mixture of gamma densities with 3 components. This distribution and its corresponding Lorenz curve were estimated using Canadian income data by Chotikapanich & Griffiths (2008), a study where survey weights could not be used. While we use the same mixture of three gamma densities, we will use 2009 household disposable income data and survey weights from the HILDA survey. This survey is a national longitudinal survey, which began in Australia in 2001 and is conducted annually (Wooden, Freidin & Watson 2002). It was initiated and funded by the Australian Government through the Department of Families, Housing, Community Services, and Indigenous Affairs, and is designed, managed, and maintained by the Melbourne Institute of Applied Economic and Social Research, University of Melbourne. The survey is a broad economic and social survey that collects key variables concerning family and household structure, as well as data on education, income, health, life satisfaction and other measures of economic and subjective wellbeing. The households are sampled using a multistage sampling design; the sampling weights are provided.

Results for standard MCMC inference (referred to as UBE) were obtained using an MCMC sample of 11000 of which 1000 were discarded as a burn in. Weighted Bayesian estimators based on using Algorithm 3 were also obtained using both the empirical distribution and the weighted finite population Bayesian bootstrap. In both cases, we generate $J = 200$ pseudo representative samples and for each PRS, we obtain a total of 5500 draws, with the first 500 draws discarded as burn in. The results were almost identical with respect to the mechanism used for generating a pseudo representative sample; for brevity, we report only the results using the WFPBB here and refer to it simply as the BWE.

All parameters for both the UBE and BWE showed evidence of convergence. The posterior means and standard deviations are reported in Table 10. The posterior means from UBE and BWE are similar in magnitude with the exception of μ_1 where there is a marked difference. The posterior standard deviations for BWE are larger, in line with the results of our Monte Carlo experiment. In Figure 3, we plot the weighted histogram, and the density estimates at the posterior means of UBE and BWE. One major difference between the two density estimates is in their ability to capture the first mode. The weighted gamma mixture fits the first mode well, but the unweighted gamma mixture overestimates the height of the density at the mode. More generally, relative to the estimates that take weights into account, the standard Bayesian estimates overstate the proportion of the population in the lower portion of the distribution, and understate the proportion of the population in the upper portion of the distribution.

Table 10. Posterior summary statistics for the parameters of individual disposable income 2009 (posterior standard deviation in brackets).

	ζ_1	ζ_2	μ_1	μ_2	μ_3	v_1	v_2	v_3
BWE	0.0565 (0.0077)	0.9106 (0.0100)	753.1687 (109.1041)	751.0602 (8.9884)	163.4528 (3.0720)	0.2295 (0.0263)	2.7266 (0.1003)	94.9858 (35.8646)
UBE	0.0571 (0.0051)	0.8999 (0.0071)	630.36 (73.7848)	723.43 (6.2133)	164.81 (1.8207)	0.2120 (0.0161)	2.6102 (0.0630)	90.1537 (18.5516)

Table 11. Posterior summary statistics of mean income, Gini, and headcount for 2009 (95% credible intervals in brackets).

	UBE	BWE
μ ($'00)	694.09 (681.75,706.93)	731.88 (714.44,750.04)
G	0.3828 (0.3758,0.3905)	0.3751 (0.3643,0.3862)
HC	0.1380 (0.1306,0.1456)	0.1169 (0.1069,0.1268)

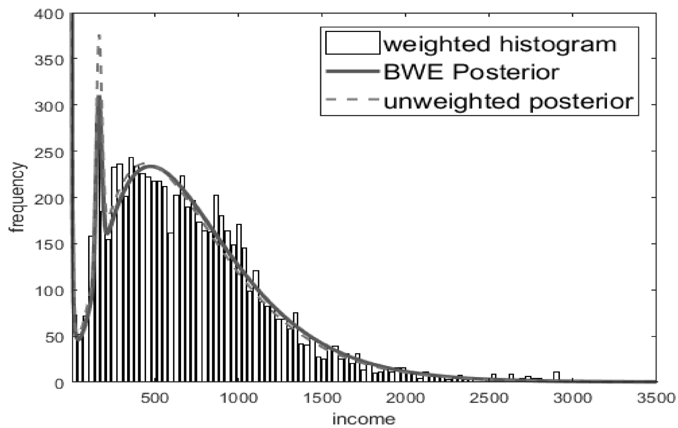

Figure 3. Weighted histogram and unweighted and weighted gamma mixture densities (at posterior means of parameters) for Australian household disposable income in 2009 ($'00).

The different estimates of the distribution have implications for three important summary statistics that are often of interest when estimating income distributions, namely mean income μ, the Gini coefficient as a measure of inequality, G, and the proportion of the population below a poverty line (the headcount ratio H). Draws from the posterior distributions of these quantities can be obtained from the following equations.

$$G = -1 + \frac{2}{\mu} \sum_{k=1}^{3} \sum_{j=1}^{3} \xi_k \xi_j \mu_k F_B \left(x_{k,j}; \nu_j, \nu_{k+1} \right) ,$$

$$H = F_G(y_p) ,$$

where $F_B \left(x_{k,j}; \nu_j, \nu_{k+1} \right)$ is the distribution function for a standard beta random variable with parameters ν_j and ν_{k+1} evaluated at $x_{k,j} = (\mu_k/\nu_k) / ((\mu_k/\nu_k) + (\mu_j/\nu_j))$ and $F_G(y_p)$ is the distribution function for the gamma mixture evaluated at a poverty line of y_p =$20000. The expression for the Gini coefficient for a mixture of gamma densities has been derived by Griffiths and Hajargasht and is available from the corresponding author on request. The posterior means and 95% credible intervals for μ, G and H are reported in Table 12. Because the distribution that ignores the weights has led to a larger estimate for the proportion of the population in the lower portion of the distribution, the unweighted estimate for μ is smaller and that for H is larger than their respective estimates from the weighted distribution. Moreover, the interval estimates for μ and H do not overlap, implying quite distinct estimates for these quantities. The difference in estimates for the Gini coefficient is less pronounced, with the unweighted estimate suggesting greater inequality.

Table 12. Posterior summary statistics of mean income, Gini, and headcount for 2009 (95% credible intervals in brackets).

	UBE	BWE
μ (\$′00)	694.09 (681.75,706.93)	730.92 (712.21,749.54)
G	0.3828 (0.3758,0.3905)	0.3759 (0.3650,0.3857)
HC	0.1380 (0.1306,0.1456)	0.1169 (0.1068,0.1278)

6. Conclusions

Empirical work in model-based inference often ignores sampling weights or makes use of the classical pseudo maximum likelihood estimator. In this paper we propose two Bayesian alternatives. Both theoretical and empirical results support the use of Bayesian weighted estimation based on the generation of a representative sample as a latent variable that can be integrated with an MCMC or other simulation algorithm. We compare methods using two Monte Carlo simulations, one using a simple Gaussian model and one with a more complex mixture of gamma densities. These simulations show that the Bayesian weighted estimator has a posterior variance that is comparable to that of the sandwich covariance matrix of the pseudo maximum likelihood estimator. This result is particularly pronounced when the weighted finite population Bayesian bootstrap is used as a scheme for simulating a pseudo representative sample. Also, using the pseudo likelihood within a Bayesian framework can lead to a posterior variance that understates the repeated sampling variation of the posterior mean, a result in line with the asymptotic theory that we have derived. An additional advantage of the Bayesian weighted estimator over the pseudo maximum likelihood estimator is that it can easily be applied to a general set of possibly complex models that can be estimated by MCMC. In an application to estimation of an Australian income distribution, we illustrate how to estimate the parameters of a three component gamma mixture model, and how to obtain posterior densities for economic quantities of interest that depend on those parameters. We find that inference about the quantities of interest, the mean income, the Gini coefficient and the headcount ratio, can be sensitive to exclusion or inclusion of the weights in the analysis.

Appendix

Consistency of BPPE

Under some regularity conditions, Walker (1969) derived the asymptotic behavior of proper posterior distributions under unweighted, independent, and identically distributed observations. Gelman et al. (2013) and Le Cam & Yang (2012) provide reviews of this area. Our results for the pseudo posterior follow a similar approach. For convenience of exposition, we assume a scalar θ but the generalisation to vector valued parameters is easily made. Let $\boldsymbol{y}$ be an $n \times 1$ random vector of finite population observations. Some aspect of the distribution of $\boldsymbol{y}$ depends on a parameter θ contained in a parameter space Θ. Assume that Θ is a closed set of points on the real line. Also assume that θ_0 is the true parameter and unique solution to the population maximization problem $\theta_0 = \max_{\theta_0 \in \Theta} \mathrm{E}_y \left[\log p(y|\theta)\right]$. For a random observed sample of size n, $y_i; i = 1, 2, \ldots, n$ we also draw I_i which is a binary indicator variable that is equal to 1 if the observation i is used in estimation. The observation y_i is observed if and only if $I_i = 1$ The sampling weights are defined as the inverse of probability of inclusion $w_i = 1/\pi_i$. Let π_i be the probability that unit i is in the sample, conditional on demographic characteristics $\boldsymbol{D}_i$ that is, $\pi_i = \Pr(I_i = 1|\boldsymbol{D} = \boldsymbol{D}_i)$.

Given the data $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$ and the sampling weights $\boldsymbol{w} = (w_1, w_2, \ldots, w_n)^\top$ and provided that the prior density $p(\theta)$ is continuous and positive, the pseudo posterior distribution can be written as:

$$\tilde{p}\left(\theta|\boldsymbol{y}, \boldsymbol{w}\right) \propto \prod_{i=1}^{n} p(y_i|\theta)^{I_i w_i} p(\theta)\,.$$

Taking logs and dividing by n gives

$$\frac{1}{n} \log \tilde{p}\left(\theta|\boldsymbol{y}, \boldsymbol{w}\right) = \frac{1}{n} \sum_{i=1}^{n} I_i w_i \log p(y_i|\theta) + \frac{1}{n} \log p(\theta) + \text{Constant}\,.$$

Let $\hat{\theta}$ be the posterior mode defined as:

$$\hat{\theta} = \max_{\theta \in \Theta} \left(\frac{1}{n} \sum_{i=1}^{n} I_i w_i \log p(y_i|\theta) + \frac{1}{n} \log p(\theta) \right)\,.$$

As $n \to \infty$ the influence of the prior diminishes and the pseudo posterior is dominated by the influence of the pseudo likelihood. Given the prior $p(\theta)$ is non-zero at $\theta = \theta_0$,

$n^{-1}\log p(\theta) \to \mathbf{0}$, and by the usual weak law of large numbers

$$\frac{1}{n}\sum_{i=1}^{n} I_i w_i \log p(y_i|\theta) = \mathrm{E}\left[\frac{I_i}{\pi_i}\log p(y_i|\theta)\right].$$

By using the law of iterated expectations we have

$$\begin{aligned}
\mathrm{E}_y\left[\frac{I_i}{\pi_i}\log p(y_i|\theta)\right] &= \int\int\int\left[\frac{I_i}{\pi_i}\log p(y_i|\theta)\right] p(y, I, \boldsymbol{D})dydId\boldsymbol{D} \\
&= \int\int\left[\frac{\int I_i p(I|y,\boldsymbol{D})dI}{\pi_i}\log p(y_i|\theta)\right] p(y,\boldsymbol{D})dyd\boldsymbol{D} \\
&= \int\int\left[\frac{\pi_i}{\pi_i}\log p(y_i|\theta)\right] p(y,\boldsymbol{D})dyd\boldsymbol{D} \\
&= \int \log p(y_i|\theta)p(y)dy \int p(\boldsymbol{D}|y)d\boldsymbol{D} \\
&= \mathrm{E}\log p(y_i|\theta).
\end{aligned}$$

where the third equality follows from $E(I_i|y_i,\boldsymbol{D}_i) = \Pr(I_i = 1|\boldsymbol{D} = \boldsymbol{D}_i) = \pi_i$. Because θ_0 is assumed to uniquely maximise $\mathrm{E}_y\left[\log p(y_i|\theta)\right]$ from assumption 1 we have $\text{plim}_{n\to\infty}\hat{\theta} = \theta_0$

Asymptotic normality of BPPE

Let $N_{\hat{\theta}}(\epsilon) = \left\{\theta : \left|\theta - \hat{\theta}\right| < \epsilon/\sqrt{n}\right\}$ be a neighbourhood of $\hat{\theta}$ contained in Θ, where $\epsilon > 0$ is a given fixed number. Using Taylor's theorem to expand $\log\tilde{p}\left(\theta|\boldsymbol{y},\boldsymbol{w}\right)$ around θ leads to

$$\begin{aligned}
\log\tilde{p}\left(\theta|\boldsymbol{y},\boldsymbol{w}\right) \approx \log\tilde{p}\left(\hat{\theta}|\boldsymbol{y},\boldsymbol{w}\right) &+ \left(\theta-\hat{\theta}\right)\left.\frac{\partial\log\tilde{p}\left(\theta|\boldsymbol{y},\boldsymbol{w}\right)}{\partial\theta}\right|_{\theta=\hat{\theta}} \\
&+ \frac{1}{2}\left(\theta-\hat{\theta}\right)^2\left.\frac{\partial^2\log\tilde{p}\left(\theta|\boldsymbol{y},\boldsymbol{w}\right)}{\partial\theta^2}\right|_{\theta=\hat{\theta}} + R,
\end{aligned}$$

where R is of higher order than $(\theta-\hat{\theta})^2$ and the term $\partial\log\tilde{p}(\theta|\boldsymbol{y},\boldsymbol{w})/\partial\theta|_{\theta=\hat{\theta}}$ is zero since the log posterior density function has zero first derivative at the posterior mode. The first term can be treated as constant since it does not involve θ. We can say that as $n\to\infty$, any θ in $N_{\hat{\theta}}(\epsilon)$ will approach $\hat{\theta}$ in probability. Thus for any small $\delta > 0$

$$\underset{n\to\infty}{lim}\Pr\left[\underset{\theta\in N_{\hat{\theta}}(\epsilon)}{sup}|R| < \delta\right] = 1.$$

In the neighbourhood $N_{\hat{\theta}}(\epsilon)$, we can express the pseudo posterior $\log \tilde{p}(\theta|\boldsymbol{y}, \boldsymbol{w})$ as follows as $n \to \infty$:

$$\tilde{p}\left(\theta|\boldsymbol{y}, \boldsymbol{w}\right) \propto \exp\left\{-\frac{n}{2}\left(\theta-\hat{\theta}\right)^2\left[-\frac{1}{n}\left.\frac{\partial^2 \log \tilde{p}\left(\theta|\boldsymbol{y}, \boldsymbol{w}\right)}{\partial\theta^2}\right|_{\theta=\hat{\theta}}\right]\right\}.$$

Now,

$$-\frac{1}{n}\left.\frac{\partial^2 \log \tilde{p}\left(\theta|\boldsymbol{y}, \boldsymbol{w}\right)}{\partial\theta^2}\right|_{\theta=\hat{\theta}} = -\frac{1}{n}\left.\frac{\partial^2 \log \tilde{p}\left(\theta\right)}{\partial\theta^2}\right|_{\theta=\hat{\theta}} - \frac{1}{n}\sum_{i=1}^{n} w_i \left.\frac{\partial^2 \log \tilde{p}\left(y_i|\theta\right)}{\partial\theta^2}\right|_{\theta=\hat{\theta}}$$

As $n \to \infty$ the first term

$$-\frac{1}{n}\left|\frac{\partial^2 \log p(\theta)}{\partial\theta^2}\right|_{\theta=\hat{\theta}}$$

goes to zero and the second term

$$-\frac{1}{n}\sum_{i=1}^{n} w_i \left.\frac{\partial^2 \log \tilde{p}\left(y_i|\theta\right)}{\partial\theta^2}\right|_{\theta=\hat{\theta}},$$

is the estimated weighted Hessian matrix evaluated at $\theta = \hat{\theta}$. Therefore as $n \to \infty$, $\tilde{p}(\theta|\boldsymbol{y}, \boldsymbol{w})$ converges to a normal distribution with mean $\hat{\theta}$ and variance

$$\sigma^2_{BPPE} = \frac{1}{n}\left(-\frac{1}{n}\sum_{i=1}^{n} w_i \left.\frac{\partial^2 \log p\left(y_i|\theta\right)}{\partial\theta^2}\right|_{\theta=\hat{\theta}}\right)^{-1}$$

in the neighbourhood of $N_{\hat{\theta}}(\epsilon)$. The next step is to ensure that θ_0 is in the neighbourhood of $\hat{\theta}$ which follows from the consistency of $\hat{\theta}$. Also, given the symmetry of the asymptotic distribution, the posterior mean will similarly have a large sample variance given by σ^2_{BPPE}.

References

AITKIN, M. (2008). Applications of the Bayesian bootstrap in finite population inference. *Journal of Official Statistics* **24**, 21–51.

ALBERT, J.H. & CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669– 679.

BREUNIG, R.V. (2001). Density estimation for clustered data. *Econometric Reviews* **20**, 353–367.

CHIB, S. (1992). Bayes inference in the Tobit censored regression model. *Journal of Econometrics* **51**, 79–99.

CHOTIKAPANICH, D. & GRIFFITHS, W. (2008). Estimating income distributions using a mixture of gamma densities. In *Modeling Income Distributions and Lorenz Curves*, ed. D. Chotikapanich. New York: Springer.

COHEN, M. (1997). The Bayesian bootstrap and multiple imputation for unequal probability sample designs. In *Proceedings of the Survey Research Method Section, American Statistical Association*. August 1997, pp. 635–638.

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

Dong, Q., Elliott, M.R. & Raghunathan, T.E. (2014b). Combining information from multiple complex surveys. *Survey Methodology* **40**, 347–354.

Gelman, A., Stern, H.S., Carlin, J.B., Dunson, D.B., Vehtari, A. & Rubin, D.B. (2013). *Bayesian Data Analysis.* Boca Raton: Chapman and Hall/CRC.

Geweke, J. (2007). Interpretation and inference in mixture models: Simple MCMC works. *Computational Statistics and Data Analysis* **51**, 3529–3550.

Geweke, J. & Amisano, G. (2011). Hierarchical Markov normal mixture models with applications to financial asset returns. *Journal of Applied Econometrics* **26**, 1–29.

Geweke, J. & Keane, M. (2007). Smoothly mixing regressions. *Journal of Econometrics* **138**, 252–290.

Godambe, V.P. & Thompson, M. (1986). Parameters of superpopulation and survey population : Their relationships and estimation. *International Statistical Review* **54**, 127–138.

Hesketh, S. & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society, Series A* **169**, 805–827.

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

Korinek, A., Mistiaen, J.A. & Ravallion, M. (2007). An econometric method of correcting for unit nonresponse bias in surveys. *Journal of Econometrics* **136**, 213–235.

Lazar, N. (1989). Bayesian empirical likelihood. *Biometrika* **90**, 319–326.

Le Cam, L. & Yang, G.L. (2012). *Asymptotics in statistics: some basic concepts.* New York: Springer Science & Business Media.

Lo, A.Y. (1993). A Bayesian method for weighted sampling. *The Annals of Statistics* **21**, 2138–2148.

Molina, E. & Skinner, C. (1992). Pseudo-likelihood and quasi-likelihood estimation for complex sampling schemes. *Computational Statistics and Data Analysis* **13**, 395–505.

Monahan, J. & Boos, D. (1992). Proper likelihoods for Bayesian analysis. *Biometrika* **79**, 271–278.

Pfeffermann, H. (1996). The use of sampling weights for survey data analysis. *Statistical Methods in Medical Research* **5**, 239–261.

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

Rubin, B.Y.D.B. (1981). The Bayesian bootstrap. *The Annals of Statistics* **9**, 130–134.

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

Schennach, S. (2005). Bayesian exponentially tilted empirical likelihood. *Biometrika* **92**, 31–46.

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

Skinner, C. & Mason, B. (2012). Weighting in the regression analysis of survey data with a cross-national application. *Canadian Journal of Statistics* **40**, 697–711.

Tanner, M.A. & Wong, W.H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82**, 528–540.

Walker, A. (1969). On the asymptotic behaviour of posterior distributions. *Journal of the Royal Statistical Society, Series B* **31**, 80–88.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48**, 817–838.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.

Wiper, M., Insua, D.R. & Ruggeri, F. (2001). Mixtures of gamma distributions with applications. *Journal of Computational and Graphical Statistics* **10**, 440–454.

WOODEN, M., FREIDIN, S. & WATSON, N. (2002). The household, income and labour dynamics in Australia (HILDA) survey: Wave 1. *Australian Economic Review* **35**, 339–348.

WOOLDRIDGE, J.M. (1999). Asymptotic properties of weighted M-estimators for variable probability samples. *Econometrica* **67**, 1385–1406.

WOOLDRIDGE, J.M. (2001). Asymptotic properties of weighted M-estimators for standard stratified samples. *Econometric Theory* **17**, 451–470.

WOOLDRIDGE, J.M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics* **141**, 1281–1301.