Bayesian Analysis (2020) **15**, Number 2, pp. 363–388

Bayesian Bootstraps for Massive Data

Andrés F. Barrientos[*] and Víctor Peña[†]

Abstract. In this article, we present data-subsetting algorithms that allow for the approximate and scalable implementation of the Bayesian bootstrap. They are analogous to two existing algorithms in the frequentist literature: the bag of little bootstraps (Kleiner et al., 2014) and the subsampled double bootstrap (Sengupta et al., 2016). Our algorithms have appealing theoretical and computational properties that are comparable to those of their frequentist counterparts. Additionally, we provide a strategy for performing lossless inference for a class of functionals of the Bayesian bootstrap and briefly introduce extensions to the Dirichlet Process.

Keywords: bootstrap, big data, Bayesian nonparametric, scalable inference.

1 Introduction

Massive datasets are increasingly common in statistical applications, mainly because current computational technologies are capable of efficiently recording and storing large datasets. As a consequence, there is an increasing demand for statistical methods that can analyze large volumes of data. Quite often, a single computer cannot store big datasets into its internal memory, and statistical analyses can only be performed in smaller subsets of the original data. In such cases, we must use algorithms that combine statistical outputs from subsets to approximate the results we would obtain if we analyzed the full dataset at once.

In the frequentist literature, two scalable adaptations of the bootstrap have been proposed: the bag of little bootstraps (BLB; Kleiner et al., 2014) and the subsampled double bootstrap (SDB; Sengupta et al., 2016). These adaptations are based on data-subsetting. The BLB proceeds by splitting the data into subsets, bootstrapping within each subset, and averaging the summaries of the "little" bootstraps to get a global assessment of an estimator. To resemble the "big" bootstrap (i.e., the bootstrap based on the full dataset), the little bootstraps need to be rescaled. The rescaling is achieved by modifying the parameters of the corresponding multinomial distribution to avoid extra computational cost. The SDB is an alternative to the BLB. The SDB proceeds by drawing random subsets from the entire dataset, running a rescaled bootstrap of size one within each subset, computing the root function for each rescaled bootstrap, and finally computing a summary of the bootstrapped values to obtain a global assessment of the estimator. As shown in Kleiner et al. (2014) and Sengupta et al. (2016), the BLB and SDB are computationally efficient and provide adequate assessments of uncertainty.

[*] Andrés F. Barrientos is Postdoctoral Associate, Department of Statistical Science, Duke University, Durham, NC 27708, afb26@duke.edu

[†] Víctor Peña is Assistant Professor, Department of Information Systems and Statistics, Baruch College, New York, NY 10010

https://doi.org/10.1214/19-BA1155

In this paper, we develop data-subsetting methods for the Bayesian bootstrap (BB) that are analogous to the BLB and the SDB. We will refer to these adaptations as the bag of little Bayesian bootstraps (BLBB) and the subsampled double Bayesian bootstrap (SDBB), respectively. The BB is a nonparametric model for probability measures proposed by Rubin (1981) as a Bayesian analogue to the bootstrap (Efron, 1979). As discussed in Lyddon et al. (2019, page 1), the BB represents a useful modeling technique to perform *general Bayes updating*, which "is a way of updating prior belief distributions that does not need the construction of a global probability model." For massive datasets, the BB is an appealing procedure for two main reasons: (1) it can accommodate complex features that are inherent to big data and (2) its implementation does not rely on recursive sampling algorithms, which are usually computationally expensive (e.g., Markov Chain Monte Carlo methods). In addition to the BLBB and SDBB, we present a strategy for performing lossless inference for a class of functionals of the BB. As a natural extension, we generalize our data-subsetting strategies to Dirichlet Processes (DP).

Our methods complement the growing literature on scalable Bayesian methods. Taddy et al. (2015, 2016) advocate for the use of the BB in massive datasets and approximate the distribution of certain functionals through Taylor series expansions and empirical Bayes procedures. Unlike these approaches, our proposal is based on data-subsetting. Most data-subsetting procedures consist of three steps: (1) the dataset is divided into subsets; (2) for each subset, a rescaled posterior distribution (subposterior) resembling the full posterior (i.e., the posterior distribution obtained by conditioning on the whole dataset) is obtained; and (3) the subposteriors are combined to either approximate the full posterior or a posterior summary of interest (e.g., a posterior variance). Methods based on subposteriors can be classified into two categories: methods that rescale the prior (see e.g., Wang and Dunson, 2013; Neiswanger et al., 2014; Wang et al., 2015; Scott et al., 2016) and methods that rescale the likelihood (see e.g. Minsker et al., 2017; Srivastava et al., 2018, 2015; Li et al., 2017). Particularly, there are some similarities between the method proposed in Li et al. (2017) and the BLBB: both methods focus on summaries of the posterior (not on the full posterior itself) and use the same rule to combine them. The methodology proposed in Li et al. (2017) applies to a large class of models; however, its theoretical guarantees are proved under assumptions (e.g., Assumption 3) that are not verifiable for the BB.

The article is organized as follows. In Section 2, we introduce and define the BB. Then, we describe the BLBB, the SDBB, and a strategy to perform lossless inference. Section 2 ends with a discussion of extensions for the DP. In Section 3, we illustrate the performance of our methods in simulation studies. In Section 4.1, we apply the BLBB and SDBB to model U.S. federal employees' wages from a subset of the Office of Personnel Management's datafile. In Section 4.2, we use our methods to model whether or not households in the American Community Survey are paying for a fire/hazard/flood insurance. The paper concludes with a discussion and directions for future work.

2 Data-subsetting strategies for the Bayesian bootstrap

Throughout, we assume that the observations $\mathcal{X}_n = \{X_1, X_2, \ldots, X_n\}$, $X_i \in \mathbb{R}^p$, are independent and identically distributed given a probability measure P, which represents the data-generating mechanism. The BB is a probability model for P given the data $\mathcal{X}_n$ which admits the stochastic representation

$$P_n^{\text{BB}}(\cdot) = \sum_{i=1}^{n} W_i \delta_{X_i}(\cdot), \tag{1}$$

where $(W_1, \ldots, W_n) \sim \text{Dirichlet}(1, \ldots, 1)$. The BB defines a distribution on probability measures which bypasses the traditional prior to posterior update, and it is nonparametric in the sense that no assumptions are made on the data-generating mechanism P beyond conditional independence of the data given P (Taddy et al., 2016). Additionally, the stochastic representation above allows us to obtain draws from P_n^{BB} directly without resorting to Markov Chain Monte Carlo methods. However, when n is massive and the full dataset cannot be loaded into memory, the BB cannot be easily implemented and approximations are needed (such as the ones presented in this article).

Several papers investigate the theoretical properties and methodological uses of the BB. From a theoretical point of view, various authors have studied its first and second-order asymptotic properties (Lo, 1987; Weng, 1989), proposed extensions and variations (Lo, 1991; Kim and Lee, 2003; Ishwaran et al., 2009), and provided distributional characterizations of functionals of P_n^{BB} (Gasparini, 1995; Choudhuri, 1998; Cifarelli and Melilli, 2000). The connections between the BB and other processes have also been an object of study. In relation to Dirichlet processes, the BB is a limiting case of the posterior distribution of a DP (as the concentration parameter of the DP goes to zero, it converges weakly to the BB), so we can interpret it as a “noninformative” limit of the DP (Lo, 1987; Muliere and Secchi, 1996; Gasparini, 1995).

There are numerous applied and methodological contributions that use the BB as a building block. For example, it has been used in areas such as censored data (Lo, 1993; James, 1997), finite population (Lo, 1988), quantile regression (Hahn, 1997), quantile estimation (Meeden, 1993), multivariate regression (Heckelei and Mittelhammer, 2003), receiver operating characteristic curve estimation (Gu et al., 2008), predictive modeling (Clyde and Lee, 2001; Fushiki, 2010), synthetic data (Dong et al., 2014), tree-based modeling (Taddy et al., 2015), high-dimensional inverse covariance matrix estimation (Datta and Ghosh, 2014), causal studies (Graham et al., 2016), multiple imputation (Rubin and Schenker, 1986; Siddique and Belin, 2008; Zhou et al., 2016), among others.

In addition to the above-mentioned favorable properties, the BB also has some potentially unappealing features. For example, P_n^{BB} is almost surely discrete, and its support is limited to the observed data $\mathcal{X}_n$. At this point, we would like to stress that the goal of the article is not to provide an extensive discussion of the advantages and shortcomings of the BB, but rather to provide methodological tools to implement the model when n is massive.

Let $\pi_P(\cdot|\mathcal{X}_n)$ be the posterior distribution of P given $\mathcal{X}_n$, which has a stochastic representation as in Expression (1). We assume that the goal is to make posterior inferences about a functional of P denote by ϕ. We use the notation $\pi_\phi(\cdot|\mathcal{X}_n)$ for the posterior distribution of $\phi(P)$. In the following subsections, we introduce methods for approximating $\xi\{\pi_\phi(\cdot|\mathcal{X}_n)\}$, where ξ is a summary of interest (e.g., mean, variance, length of credible intervals).

2.1 Bag of little Bayesian bootstraps

The BLBB is an adaptation of the bag of little bootstraps proposed by Kleiner et al. (2014). This procedure provides an approximation of $\xi\{\pi_\phi(\cdot|\mathcal{X}_n)\}$ when n is large, and it comprises three steps: divide, rescale, and combine. In the first step, we randomly split the dataset into subsets of size b such that each subset can be stored in the internal memory of the computer. We define these subsets by generating a random partition $\mathcal{I}_{1,b,n},\dots,\mathcal{I}_{n/b,b,n}$ of the set $\{1,\dots,n\}$, where $|\mathcal{I}_{j,b,n}| = b$. For ease of exposition, we assume that n is a multiple of b. In the second step, we compute a rescaled version of the posterior distribution of $\phi(P)$ for each dataset $\mathcal{X}_{j,b,n} = \{X_i\}_{i\in\mathcal{I}_{j,b,n}}$, $j = 1,\dots,n/b$. In the third step, we combine the summaries found with the rescaled posteriors to obtain an approximation of $\xi\{\pi_\phi(\cdot|\mathcal{X}_n)\}$.

The purpose of rescaling is to define a posterior distribution that resembles the one we would obtain with the full dataset. Without rescaling, the next step after partitioning the dataset would be to compute $\xi\{\pi_\phi(\cdot|\mathcal{X}_{j,b,n})\}$, $j = 1,\dots,n/b$, and then combine these summaries using, for example, the average; that is, to approximate $\xi\{\pi_\phi(\cdot|\mathcal{X}_n)\}$ by $b/n\sum_{j=1}^{n/b}\xi\{\pi_\phi(\cdot|\mathcal{X}_{j,b,n})\}$. This strategy could work if the $\xi\{\pi_\phi(\cdot|\mathcal{X}_{j,b,n})\}$ provided a reasonable approximation of $\xi\{\pi_\phi(\cdot|\mathcal{X}_n)\}$. Unfortunately, this is often not the case. For example, if ξ is the variance of $\pi_\phi(\cdot|\mathcal{X}_n)$, then we expect $\xi\{\pi_\phi(\cdot|\mathcal{X}_n)\} < \xi\{\pi_\phi(\cdot|\mathcal{X}_{j,b,n})\}$, which in turn implies that $\xi\{\pi_\phi(\cdot|\mathcal{X}_n)\} < b/n\sum_{j=1}^{n/b}\xi\{\pi_\phi(\cdot|\mathcal{X}_{j,b,n})\}$. For this reason, we rescale the posterior distributions within each subset by following a standard strategy employed in several data-splitting based procedures (e.g., Kleiner et al., 2014; Minsker et al., 2017; Srivastava et al., 2015). Each subset $\mathcal{X}_{j,b,n}$ is replicated n/b times such that the replicated dataset contains n instead of b data points. Then, we obtain the posterior distribution associated with the replicated dataset, which we refer to as the "rescaled posterior distribution."

For a functional $\phi(P)$, the rescaled posterior distribution of $\phi(P)$ given $\mathcal{X}_{j,b,n}$ is defined as $\pi_\phi^{\text{BLBB}}(\cdot|\mathcal{X}_{j,b,n}) = \pi_\phi(\cdot|\mathcal{X}^*_{j,b,n})$, where $\mathcal{X}^*_{j,b,n}$ denotes the dataset $\mathcal{X}_{j,b,n}$ replicated n/b times, so we have a new sample of size n. Note that rescaling $\pi_\phi^{\text{BLBB}}(\cdot|\mathcal{X}_{j,b,n})$ can be achieved by rescaling the posterior distribution of P given $\mathcal{X}_{j,b,n}$. The rescaled posterior distribution of P given $\mathcal{X}_{j,b,n}$ is denoted $\pi_P^{\text{BLBB}}(\cdot|\mathcal{X}_{j,b,n}) = \pi_P(\cdot|\mathcal{X}^*_{j,b,n})$. In this case, the distribution $\pi_P^{\text{BLBB}}(\cdot|\mathcal{X}_{j,b,n})$ also has a stochastic representation of the form

$$P_{j,b,n}^{\text{BLBB}}(\cdot) = \sum_{i\in\mathcal{I}^*_{j,b,n}} W_{i,j}\delta_{X_i}(\cdot) \overset{d}{=} \sum_{i\in\mathcal{I}_{j,b,n}} W^*_{i,j}\delta_{X_i}(\cdot), \tag{2}$$

where $(W_{i,j})_{i\in\mathcal{I}^*_{j,b,n}} \sim \text{Dirichlet}(1,\dots,1)$, $(W^*_{i,j})_{i\in\mathcal{I}_{j,b,n}} \sim \text{Dirichlet}(n/b,\dots,n/b)$, $\overset{d}{=}$ denotes equality in distribution, and $\mathcal{I}^*_{j,b,n}$ denotes the subset $\mathcal{I}_{j,b,n}$ replicated n/b times. The process in Expression (2) belongs to the class of BB clones proposed in Lo (1991). With this stochastic representation, we have $\phi(P_{j,b,n}^{\text{BLBB}})|\mathcal{X}_{j,b,n} \sim \pi_\phi^{\text{BLBB}}(\cdot|\mathcal{X}_{j,b,n})$. Although we compute the rescaled posterior using a replicated dataset of size n, the computational cost of drawing from $\pi_\phi^{\mathrm{BLBB}}(\cdot|\mathcal{X}_{j,b,n})$ is the same as of a BB with sample size equal to b. Thus, we propose approximating $\xi\{\pi_\phi(\cdot|\mathcal{X}_n)\}$ by $b/n\sum_{j=1}^{n/b}\xi\{\pi_\phi^{\mathrm{BLBB}}(\cdot|\mathcal{X}_{j,b,n})\}$.

We provide theoretical guarantees for approximating summaries of $\phi_T(P_n^{\mathrm{BB}}, \mathbb{P}_n) = \sqrt{n}(T(P_n^{\mathrm{BB}}) - T(\mathbb{P}_n))$ by summaries of $\phi_T(P_{j,b,n}^{\mathrm{BLBB}}, \mathbb{P}_{j,b,n}) = \sqrt{n}(T(P_{j,b,n}^{\mathrm{BLBB}}) - T(\mathbb{P}_{j,b,n}))$, where T is a functional and, $\mathbb{P}_n$ and $\mathbb{P}_{j,b,n}$ are the empirical measures associated with $\mathcal{X}_n$ and $\mathcal{X}_{j,b,n}$, respectively. Since $E[P_n^{\mathrm{BB}}|\mathcal{X}_n] = \mathbb{P}_n$ and $E[P_{j,b,n}^{\mathrm{BLBB}}|\mathcal{X}_{j,b,n}] = \mathbb{P}_{j,b,n}$, we can think of $T(\mathbb{P}_n)$ and $T(\mathbb{P}_{j,b,n})$ as measures of central tendency for the distribution of $T(P_n^{\mathrm{BB}})|\mathcal{X}_n$ and $T(P_{j,b,n}^{\mathrm{BLBB}})|\mathcal{X}_{j,b,n}$, respectively. The functional ϕ_T can be used to quantify uncertainty (e.g., estimate the length of intervals and measures of dispersion) associated with the functional T. The study of the asymptotic properties of $\phi_T(P_n^{\mathrm{BB}}, \mathbb{P}_n)$ and $\phi_T(P_{j,b,n}^{\mathrm{BLBB}}, \mathbb{P}_{j,b,n})$ requires understanding the asymptotic behavior of the processes $\sqrt{n}(P_n^{\mathrm{BB}} - \mathbb{P}_n)$ and $\sqrt{n}(P_{j,b,n}^{\mathrm{BLBB}} - \mathbb{P}_{j,b,n})$. These processes belong to the class of *weighted bootstrap empirical processes* studied in Section 3.6.2 of van der Vaart and Wellner (1996). The connection with *weighted bootstrap empirical processes* and an assumption on the differentiability of T allow us to prove that $b/n\sum_{j=1}^{n/b}\xi\{\pi_{\phi_T}^{\mathrm{BLBB}}(\cdot|\mathcal{X}_{j,b,n})\}$ suitably approximates $\xi\{\pi_{\phi_T}(\cdot|\mathcal{X}_n)\}$, where $\pi_{\phi_T}^{\mathrm{BLBB}}(\cdot|\mathcal{X}_{j,b,n})$ and $\pi_{\phi_T}(\cdot|\mathcal{X}_n)$ denote the distributions of $\phi_T(P_{j,b,n}^{\mathrm{BLBB}}, \mathbb{P}_{j,b,n})|\mathcal{X}_{j,b,n}$ and $\phi_T(P_n^{\mathrm{BB}}, \mathbb{P}_n)|\mathcal{X}_n$, respectively. The proof of this result is similar to the proof of Theorem 1 in Kleiner et al. (2014), and it also relies on the assumption that T is Hadamard differentiable at P_0 (the data generation mechanism) and the existence of a P_0-Donsker class. The specific statement of our result and its proof can be found in Theorem 1 in the supplementary material (Barrientos and Peña, 2019). We show that even if we use s instead of n/b subsets ($s < n/b$), the average $s^{-1}\sum_{j=1}^{s}\xi\{\pi_{\phi_T}^{\mathrm{BLBB}}(\cdot|\mathcal{X}_{j,b,n})\}$ can provide a reasonable approximation of $\xi\{\pi_{\phi_T}(\cdot|\mathcal{X}_n)\}$. Figure 1 describes the resulting Monte Carlo algorithm for the BLBB.

2.2 Subsampled double Bayesian bootstrap

The SDBB is the Bayesian analogue to the subsampled double bootstrap for massive data proposed by Sengupta et al. (2016), which also provides an approximation of $\xi\{\pi_\phi(\cdot|\mathcal{X}_n)\}$. In Sengupta et al. (2016), the authors claim that the SDB outperforms the BLB in some scenarios with limited time budget, especially when it is only possible to run $s < n/b$ little bootstraps. Therefore, we would expect the same phenomenon to occur with the BLBB and SDBB.

Theorem 1 in the supplementary material shows that $s^{-1}\sum_{j=1}^{s}\xi\{\pi_{\phi_T}^{\mathrm{BLBB}}(\cdot|\mathcal{X}_{j,b,n})\}$ can provide a reasonable approximation of $\xi\{\pi_{\phi_T}(\cdot|\mathcal{X}_n)\}$. However, the BLBB could be outperformed by the SDBB because the little Bayesian bootstraps only consider a unique partition of the dataset and, if the computational budget is limited, only a fraction of the dataset contributes to the analysis. We refer to this fraction as *sample coverage*, a term we borrow from Sengupta et al. (2016).

The SDBB is a procedure that ensures a higher sample coverage compared to the BLBB and does not require using a partition of the dataset. Instead, this procedure uses random subsets of $\mathcal{X}_n$. Let $\mathcal{X}_{b,n} = \{X_{R_1}, \ldots, X_{R_b}\}$ be representing the random subset,

```
Input: Data $\mathcal{X}_n = \{X_1, \dots, X_n\}$
       $\phi_T$: functional of interest
       $\xi$: posterior summary of interest
       $b$: subset size
       $s$: number of sampled subsets

Output: approximation of $\xi\left(\pi_{\phi_T}(\cdot|\mathcal{X}_n)\right)$

Generate a random partition $\mathcal{I}_{1,b,n}, \dots, \mathcal{I}_{n/b,b,n}$ of the set $\{1, \dots, n\}$, where $|\mathcal{I}_{j,b,n}| = b$

for $j \leftarrow 1$ to $s$ do
    $\mathbb{P}_{j,b,n}(\cdot) \leftarrow b^{-1} \sum_{i \in \mathcal{I}_{j,b,n}} \delta_{X_i}(\cdot)$
    for $l \leftarrow 1$ to $r$ do
        sample $(w_{i,j})_{i \in \mathcal{I}_{j,b,n}} \sim$ Dirichlet$(n/b, \dots, n/b)$
        $P^{(l)}_{j,b,n}(\cdot) \leftarrow \sum_{i \in \mathcal{I}_{j,b,n}} w_{i,j} \delta_{X_i}(\cdot)$
        $t^{(l)}_{j,b,n} \leftarrow T\left(P^{(l)}_{j,b,n}\right) - T\left(\mathbb{P}_{j,b,n}\right)$
    end
    $\pi^{\text{BLBB}}_{\phi_T}(\cdot|\mathcal{X}_{j,b,n}) \leftarrow r^{-1} \sum_{l=1}^{r} \delta_{t^{(l)}_{j,b,n}}(\cdot)$
    $\xi^{\text{BLBB}}_{j,b,n} \leftarrow \xi\left(\pi^{\text{BLBB}}_{\phi_T}(\cdot|\mathcal{X}_{j,b,n})\right)$
end
return $s^{-1} \sum_{j=1}^{s} \xi^{\text{BLBB}}_{j,b,n}$
```

```
Input: Data $\mathcal{X}_n = \{X_1, \dots, X_n\}$
       $\phi_T$: functional of interest
       $\xi$: posterior summary of interest
       $b$: subset size
       $s$: number of sampled subsets

Output: approximation of $\xi\left(\pi_{\phi_T}(\cdot|\mathcal{X}_n)\right)$

for $j \leftarrow 1$ to $s$ do
    sample $(w_1, \dots, w_b) \sim$ Dirichlet$(n/b, \dots, n/b)$
    sample $(r_1, \dots, r_b) \sim \mathrm{U}\left(\mathcal{P}^n_b\right)$
    $\mathbb{P}_{j,b,n}(\cdot) \leftarrow b^{-1} \sum_{i=1}^{b} \delta_{X_{r_i}}(\cdot)$
    $P^{(j)}_{b,n}(\cdot) \leftarrow \sum_{i=1}^{b} w_i \delta_{X_{r_i}}(\cdot)$
    $t^{(j)}_{b,n} \leftarrow T\left(P^{(j)}_{b,n}\right) - T\left(\mathbb{P}_{j,b,n}\right)$
end
$\pi^{\text{SDBB}}_{\phi_T}(\cdot|\mathcal{X}_n) \leftarrow s^{-1} \sum_{j=1}^{s} \delta_{t^{(j)}_{b,n}}(\cdot)$
return $\xi\left(\pi^{\text{SDBB}}_{\phi_T}(\cdot|\mathcal{X}_n)\right)$
```

Figure 1: Monte Carlo algorithms for the BLBB (left) and SDBB (right).

where $b \in \mathbb{N}$ is such that $\mathcal{X}_{b,n}$ can be stored in the internal memory of the computer, $R = (R_1, \dots, R_b) \sim \mathrm{U}(\mathcal{P}^n_b)$, and $\mathrm{U}(\mathcal{P}^n_b)$ stands for the uniform distribution defined on the permutations of size b of the elements $\{1, \dots, n\}$. The SDBB runs a very little Bayesian bootstrap of size 1 for each drawn subset, so it has higher sample coverage than the BLBB. The use of subsets of $\mathcal{X}_n$ also requires a rescaling strategy, and we use the same one that was used for the BLBB. We approximate the posterior distribution of P given $\mathcal{X}_n$ by $\pi^{\text{SDBB}}_P(\cdot|\mathcal{X}_n)$, where $\pi^{\text{SDBB}}_P(\cdot|\mathcal{X}_n)$ is the distribution induced by the SDBB process. We define the SDBB process as

$$P^{\text{SDBB}}_{b,n}(\cdot) = \sum_{i \in \mathcal{I}^*_b} W_i \delta_{X_{R_i}}(\cdot) \overset{d}{=} \sum_{i=1}^{b} W^*_i \delta_{X_{R_i}}(\cdot),$$

where $(W_i)_{i \in \mathcal{I}^*_b} \sim \text{Dirichlet}(1, \dots, 1)$, $(W^*_i)_{i=1}^{b} \sim \text{Dirichlet}(n/b, \dots, n/b)$, and $\mathcal{I}^*_b$ denotes the subset $\{1, \dots, b\}$ replicated n/b times. Let $\pi^{\text{SDBB}}_{\phi}(\cdot|\mathcal{X}_n)$ be the distribution of $\phi(P^{\text{SDBB}}_{b,n})|\mathcal{X}_n$. Our proposal is to approximate $\xi\{\pi_\phi(\cdot|\mathcal{X}_n)\}$ by $\xi\{\pi^{\text{SDBB}}_\phi(\cdot|\mathcal{X}_n)\}$. We provide theoretical support for functionals of the form $\phi_T(P^{\text{BB}}_n, E[P^{\text{BB}}_n|\mathcal{X}_n]) = \sqrt{n}(T(P^{\text{BB}}_n) - T(\mathbb{P}_n))$ (this is similar to the theoretical results in Section 2.1), where T is a functional conditions under which $\xi\{\pi^{\text{SDBB}}_{\phi_T}(\cdot|\mathcal{X}_n)\}$ approximates $\xi\{\pi_{\phi_T}(\cdot|\mathcal{X}_n)\}$, where $\pi^{\text{SDBB}}_{\phi_T}(\cdot|\mathcal{X}_n)$ is the distribution of the functional $\phi_T(P^{\text{SDBB}}_{b,n}, E[P^{\text{SDBB}}_{b,n}|\mathcal{X}_n, R]) = \sqrt{n}(T(P^{\text{SDBB}}_{b,n}) - (E[P^{\text{SDBB}}_{b,n}|\mathcal{X}_n, R]))$ given $\mathcal{X}_n$, with $E[P^{\text{SDBB}}_{b,n}|\mathcal{X}_n, R] = b^{-1}\sum_{i=1}^{b} \delta_{X_{R_i}}$. The technical conditions assumed for T and P_0 in Theorem 2 are similar to those assumed for the BLBB. Theorem 2 is the counterpart of Theorem 1 in Sengupta et al. (2016). Figure 1 describes the Monte Carlo algorithm for running the SDBB.

2.3 Lossless inference for the Bayesian bootstrap

For a certain class of functionals, we can perform exact (lossless) inference for the BB after splitting the data. This strategy is based on decomposing the Dirichlet weights into Gamma random variables, and it is similar in spirit to the strategy devised for bagging in Lee and Clyde (2004). The functionals for which lossless inference can be performed are of the form $\phi(P_n^{\text{BB}}) = g\left(\int \rho(x) P_n^{\text{BB}}(dx)\right)$ where g is a measurable function, $\int \rho(x) P_n^{\text{BB}}(dx) = \left(\int \rho_1(x) P_n^{\text{BB}}(dx), \ldots, \int \rho_k(x) P_n^{\text{BB}}(dx)\right)$, $\rho = (\rho_1, \rho_2, \ldots, \rho_k)$, and ρ_l is a function defined on the sample space. This class of functionals clearly contains moments and expectations of transformations, but it also contains other functionals such as weighted least squares estimators (as we show in the example below) or the instrumental variables estimator presented in Section 2 in Chamberlain and Imbens (2003) (see the supplementary material for further details).

Example 1 (Weighted least squares)**.** *Let $Y_i \in \mathbb{R}$ be the outcome and $U_i \in \mathbb{R}^{p+1}$ be covariates, and assume that we want to model $E[Y_i|U_i]$ using a linear combination of the predictors. If the pairs are $(Y_i, U_i)|P \overset{\text{iid}}{\sim} P$ and P given the data is the BB, then we can define the least squares functional (which, in this case, has a posterior distribution):*

$$\phi(P_n^{\text{BB}}) = \boldsymbol{\beta}_{\text{lm},n}^{\text{BB}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} W_i (Y_i - U_i^\top \boldsymbol{\beta})^2.$$

This functional has been used, for instance, in Clyde and Lee (2001) and Taddy et al. (2016). We can rewrite

$$\begin{aligned} \boldsymbol{\beta}_{\text{lm},n}^{\text{BB}} &= \left(\sum_{i=1}^{n} W_i U_i U_i^\top \right)^{-1} \left(\sum_{i=1}^{n} W_i U_i Y_i \right), \\ &= \left(\int \rho_1(x) P_n^{\text{BB}}(dx) \right)^{-1} \left(\int \rho_2(x) P_n^{\text{BB}}(dx) \right), \\ &= g\left(\int \rho(x) P_n^{\text{BB}}(dx) \right), \end{aligned}$$

where $x = (y, u)$, $y \in \mathbb{R}$, $u \in \mathbb{R}^{p+1}$, $\rho_1(x) = uu^\top$, $\rho_2(x) = uy$, $g(M, v) = M^{-1}v$, M is a $(p+1) \times (p+1)$-dimensional matrix, and v is a $(p+1)$-dimensional vector. Thus, $\boldsymbol{\beta}_{\text{lm},n}^{\text{BB}}$ is included in the class of functionals for which lossless inference can be performed.

The algorithm requires processing all the n/b subsets, so it can be significantly slower than the BLBB or SDBB. A figure that summarizes the Monte Carlo algorithm for performing lossless inference can be found in the supplementary material.

2.4 Extension to the Dirichlet Process

In the subsection, we extend the results in previous subsections for the Dirichlet process (DP), which we denote $X_i|P \overset{\text{iid}}{\sim} P$, $i = 1, \ldots, n$, with $P \sim \text{DP}(\alpha, H)$. The hyperparameters of the DP are the base measure H and the concentration parameter $\alpha > 0$. The base measure H is the prior expectation of P, whereas α controls how concentrated P

is around H. The DP has the following explicit stochastic representation (Sethuraman, 1994)

$$P^{\text{DP}}(\cdot) = \sum_{k=1}^{\infty} V_k \prod_{l=1}^{k-1}(1-V_l)\,\delta_{A_k}(\cdot), \quad V_k \overset{\text{iid}}{\sim} \text{Beta}(1,\alpha), \quad A_k \overset{\text{iid}}{\sim} H,$$

and it is a conjugate model:

$$P|\mathcal{X}_n \sim \text{DP}\left(\alpha+n, \frac{\alpha}{\alpha+n}H + \frac{n}{\alpha+n}\mathbb{P}_n\right),$$

where $\mathbb{P}_n$ is the empirical probability measure. Pitman (1996) shows that the posterior above can be represented as

$$P_n^{\text{DP}}(\cdot) = R_n P^{\text{DP}}(\cdot) + (1-R_n)P_n^{\text{BB}}(\cdot), \; R_n \sim \text{Beta}(\alpha, n). \tag{3}$$

The random weight R_n, the prior measure P^{DP} of the DP, and the BB P_n^{BB} are independent given the data. For large n, the posterior measure of the DP is very close to the BB: if the same random P_n^{BB} is to be used for an analysis with the BB and the DP,

$$\mathbb{P}[d_{\text{TV}}(P_n^{\text{BB}}, P_n^{\text{DP}}) > \epsilon] = \mathbb{P}[R_n \sup_A |P_n^{\text{BB}}(A) - P_n^{\text{DP}}(A)| > \epsilon] \le \mathbb{P}[R_n \ge \epsilon].$$

For instance, if $0 < \epsilon < 1$ and $0 < \alpha \le 1$, the inequality above implies that $\mathbb{P}[d_{\text{TV}}(P_n^{\text{BB}}, P_n^{\text{DP}}) > \epsilon] \le (1-\epsilon)^{(n+1)}$.

From a more theoretical perspective, the representation in Equation (3) allows us to define an analogue of the BLBB and SDBB

$$P_{j,b,n}^{\text{BLDP}} = R_n P^{\text{DP}}(\cdot) + (1-R_n)P_{j,b,n}^{\text{BLBB}}(\cdot) \text{ and } P_{b,n}^{\text{SDDP}} = R_n P^{\text{DP}}(\cdot) + (1-R_n)P_{b,n}^{\text{SDBB}}(\cdot),$$

where BLDP and SDDP stand for bag of little Dirichlet processes and subsampled double Dirichlet process, respectively, and $R_n \sim \text{Beta}(\alpha, n)$. The Bernstein-von Mises results for the Dirichlet process that are available in the literature (see e.g., Lo, 1983; James, 2008; Varron, 2014; Castillo and Nickl, 2014) allow us to show that the asymptotic behavior of $\sqrt{n}(P_{j,b,n}^{\text{BLDP}} - E[P_{j,b,n}^{\text{BLDP}}|\mathcal{X}_{j,b,n}])$ and $\sqrt{n}(P_{b,n}^{\text{SDDP}} - E[P_{b,n}^{\text{SDDP}}|\mathcal{X}_n])$ is the same as that of $\sqrt{n}(P_n^{\text{DP}} - E[P_n^{\text{DP}}|\mathcal{X}_n])$, which is a parallel of the results found in the previous subsections (a proof can be found in the supplementary material). In order to have the same theoretical guarantees for a functional $\phi_T(P_n^{\text{BB}}, \mathbb{P}_n)$ (as described in previous subsections), the next step would be invoking the functional delta theorem. We finalize this subsection by noting that for the class of functionals that was defined in Subsection 2.3, one can perform lossless inference for the Dirichlet-Multinomial process (Kingman, 1975; Pitman, 1995), which is an approximation to the Dirichlet Process (Ishwaran and Zarepour, 2002; Muliere and Secchi, 1996). The details can be found in the supplementary material.

3 Numerical experiments

We compare the performance of the BLBB and SDBB in approximating posterior means and standard deviations, as well as the length of 95% credible interval of functionals of the BB in linear, logistic, and mixed-effects regression. The sample sizes of the datasets are always equal to 10,000. We simulate data from the following models:

- *Linear regression:* $Y_i = U_i^\top \boldsymbol{\beta}_0 + \epsilon_i$,
- *Logistic regression:* $Y_i \sim \text{Bernoulli}(p_i),\ p_i = (1 + \exp(-0.01 U_i^\top \boldsymbol{\beta}_0 + 0.25))^{-1}$,
- *Mixed-effects regression:* $Y_{ij} = \alpha_j + U_{ij}^\top \boldsymbol{\beta}_0 + \epsilon_i,\ j = 1, 2, 3$,

where $i = 1, \dots, n$, Y_i is the outcome, U_i denotes a $(p+1)$-dimensional vector containing the predictors, and $\boldsymbol{\beta}_0 = (\beta_{0,0}, \dots, \beta_{p,0})^\top = (1, 1, \dots, 1)$. The first component of U_i is equal to 1 (acting as an intercept) and the remaining elements are independent and identically distributed as a Student-t with 3 degrees of freedom. In the linear model, the errors ϵ_is are simulated from a Skew Normal distribution with location parameter -0.71, scale 1, and slant 2, which has mean 0 and is asymmetric. In the mixed model, we draw the random the effect α_j and error ϵ_i from a Skew Normal distribution with location parameter -0.71, scale 1, and slant 2 and from a Student-t with 3 degrees of freedom, respectively. Finally, in linear regression, p is taken to be 100, whereas in the case of logistic and mixed-effect regression p is equal to 25.

Throughout, we model the data as $(Y_i, U_i)|P \overset{\text{iid}}{\sim} P$ and P given the data using P_n^{BB}. We focus on the following 3 functionals that induce posterior distributions for the regression coefficients:

- *Linear regression:* the weighted least squares estimator (see, for example, Clyde and Lee, 2001 and Taddy et al., 2016),

$$\boldsymbol{\beta}_{\text{lm},n}^{\text{BB}} = \left(\sum_{i=1}^{n} W_i U_i U_i^\top\right)^{-1} \left(\sum_{i=1}^{n} W_i U_i Y\right), \tag{4}$$

- *Logistic regression:* the robust estimator, (see, for instance, Carroll and Pederson, 1993),

$$\boldsymbol{\beta}_{\text{lg},n}^{\text{BB}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\max} \sum_{i=1}^{n} W_i \left[-\log(1 + \exp(U_i^\top \boldsymbol{\beta})) + Y_i U_i^\top \boldsymbol{\beta}\right], \tag{5}$$

- *Mixed-effects regression:* the weighted estimator,

$$\boldsymbol{\beta}_{\text{mx},n}^{\text{BB}} = \left(\sum_{i=1}^{n} W_i \mathbf{U}_i^\top \hat{V} \mathbf{U}_i\right)^{-1} \left(\sum_{i=1}^{n} W_i \mathbf{U}_i^\top \hat{V} \mathbf{Y}_i\right), \tag{6}$$

where $\mathbf{U}_i = (U_{i,1}, U_{i,2}, U_{i,3})^\top$, $\mathbf{Y}_i = (Y_{i,1}, Y_{i,2}, Y_{i,3})^\top$, and $\hat{V}$ is the maximum likelihood estimator of the covariance matrix derived from the marginal likelihood of a mixed-effects model with random intercept and under Gaussian assumptions. Welsh and Richardson (1997) and Jacqmin-Gadda et al. (2007) discuss and assess the robustness of these type of functionals.

Now, we explain how the length of credible intervals, standard deviations, and mean for $\boldsymbol{\beta}^{\text{BB}}_{\text{lm},n} = T(P^{\text{BB}}_n)$ are computed, using the notation introduced in Section 2 (the summaries for $\boldsymbol{\beta}^{\text{BB}}_{\text{lg},n}$ and $\boldsymbol{\beta}^{\text{BB}}_{\text{mx},n}$ are computed using the same procedure). In the simulation studies,

$$\phi_T(P^{\text{BB}}_n, \mathbb{P}_n) = (T(P^{\text{BB}}_n) - T(\mathbb{P}_n)) = (\boldsymbol{\beta}^{\text{BB}}_{\text{lm},n} - \hat{\boldsymbol{\beta}}_{\text{lm},n}),$$

where $T(P^{\text{BB}}_n) = \boldsymbol{\beta}^{\text{BB}}_{\text{lm},n}$ and $T(\mathbb{P}_n) = \hat{\boldsymbol{\beta}}_{\text{lm},n}$. We use the notation $\xi_{l,1}$, $\xi_{l,2}$, and $\xi_{l,3}$ for the 2.5th and 97.5th percentiles, and standard deviation of the l-th marginal distribution of a $(p+1)$-dimensional distribution. Thus, we define

$$\begin{aligned}
\xi^{\text{BB}}_{l,k} &= \xi_{l,k}\{\pi_{\phi_T}(\cdot|\mathcal{X}_n)\}, \\
\xi^{\text{BLBB}}_{l,k} &= b/n \sum_{j=1}^{n/b} \xi_{l,k}\{\pi^{\text{BLBB}}_{\phi_T}(\cdot|\mathcal{X}_{j,b,n})\}, \\
\xi^{\text{SDBB}}_{l,k} &= \xi_{l,k}\{\pi^{\text{SDBB}}_{\phi_T}(\cdot|\mathcal{X}_n)\},
\end{aligned}$$

where $\xi^{\text{BLBB}}_{l,k}$ and $\xi^{\text{SDBB}}_{l,k}$ are computed using the algorithms displayed in Figure 1. With these summaries, we can compute the average relative errors of lengths of 95% credible intervals and posterior standard deviations. For example, for the BLBB, we can compute the average relative error as

$$(p+1)^{-1} \sum_{l=0}^{p} \left| (\tilde{\xi}^{\text{BLBB}}_{l,2} - \tilde{\xi}^{\text{BLBB}}_{l,1})/(\tilde{\xi}_{l,2} - \tilde{\xi}_{l,1}) - 1 \right|, \tag{7}$$

and the same computation can be carried out for the SDBB by substituting $\tilde{\xi}^{\text{BLBB}}_{l,1}$ and $\tilde{\xi}^{\text{BLBB}}_{l,2}$ by $\tilde{\xi}^{\text{SDBB}}_{l,1}$ and $\tilde{\xi}^{\text{SDBB}}_{l,2}$. For the posterior mean, we define

$$\begin{aligned}
\xi^{\text{BB}}_4 &= E[T(P^{\text{BB}}_n)|\mathcal{X}_n] = E[\boldsymbol{\beta}^{\text{BB}}_{\text{lm},n}|\mathcal{X}_n], \\
\xi^{\text{BLBB}}_4 &= b/n \sum_{j=1}^{n/b} E[T(P^{\text{BLBB}}_{j,b,n})|\mathcal{X}_{j,b,n}] = b/n \sum_{j=1}^{n/b} E[\boldsymbol{\beta}^{\text{BLBB}}_{\text{lm},n}|\mathcal{X}_{j,b,n}], \\
\xi^{\text{SDBB}}_4 &= E[T(P^{\text{SDBB}}_{b,n})|\mathcal{X}_n] = E[\boldsymbol{\beta}^{\text{SDBB}}_{\text{lm},n}|\mathcal{X}_n],
\end{aligned}$$

and quantify the bias by computing the absolute errors

$$\|\xi^{\text{BB}}_4 - \xi^{\text{BLBB}}_4\|_1 \text{ and } \|\xi^{\text{BB}}_4 - \xi^{\text{SDBB}}_4\|_1 \tag{8}$$

for the BLBB and SDBB, respectively. We use the errors in (7) and (8) to assess the performance of the methods. We compare the results found after 1,000 samples from the BB (run on the full dataset), and the results shown are averages over 100 simulated datasets. In the case of the BLBB, the number of bootstrap samples within each subgroup is equal to 100. On the other hand, the SDBB is run until 1,000 samples are drawn. For both methods, we set $b = n^{\gamma}$, $\gamma = 0.6,\ 0.7$, and 0.8.

Given that the BLBB and SDBB have asymptotic guarantees (which are proved in Theorems 1 and 2 in the supplementary material), we compare the performance of the BLBB and SDBB with two procedures based on the asymptotic distributions of $T(\mathbb{P}_n)$

and $T(\mathbb{P}_{j,b,n})$. The first procedure uses an estimate of mean and variance of the asymptotic normal distribution of $T(\mathbb{P}_n)$ to compute the summaries of interest. We refer to this method based on asymptotic normality as AN. The second one relies on data-subsetting and, for each $j = 1, \dots, n/b$, uses an estimate of mean and variance (re-scaled by a factor of b/n) of the asymptotic normal distribution of $T(\mathbb{P}_{j,b,n})$ to compute an aggregated estimation of the corresponding summaries. We aggregate estimates using the average over $j = 1, \dots, n/b$. We refer to this method as ANS. The summaries computed under AN and ANS are compared to $\xi_{l,k}^{\text{BB}}$, $k = 1, 2, 3$, and ξ_4^{BB} using (7) and (8). The estimated mean and variances of these asymptotic distributions as well as the functionals (4) and (5), and the variance $\hat{V}$ in (6) are computed using the statistical software R (R Core Team, 2015). For linear and logistic regression, we used the function `lm` and `glm` in the `stats` package; for mixed-effects regression, we used the function `lme` in library `nlme` (Pinheiro et al., 2018). The simulations were performed on a computer with Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50GHz processor and 16 GB RAM and running R version 3.3.2. We acknowledge that the results are contingent on our computing infrastructure and coding abilities. Section 3.1 below explains the results of the simulation studies and Section 3.2 contains a discussion on the computational overhead of the methods.

3.1 Monte Carlo results

In this subsection, we discuss and compare the Monte Carlo results obtained with the BB to those obtained with the BLBB, SDBB, ANS, and AN. Table 1 displays the average relative and absolute errors associated with the summaries.

In our simulations, ANS is outperformed by the BLBB and SDBB, but AN outperforms both of our methods. The good performance of AN is in agreement with our theoretical results which assert that, under appropriate regularity conditions, the posterior distribution of functionals of P_n^{BB} will be approximately normal when the sample size is large enough. However, AN requires storing the full dataset in the internal memory, which is unfeasible in realistic applications with massive data. In other words, we propose to use the BLBB and SDBB when the data cannot be stored into the internal memory of a single computer, so AN is not a direct competitor.

As expected, Table 1 shows that the performance of the proposed methods depends on the functional of interest. In linear and mixed-effects regression the errors are rather small, whereas in logistic regression the errors tend to be higher. For the functionals considered in this simulation study, we observe that the performance of the BLBB and SDBB suffers in scenarios where the functional of interest does not have a closed-form expression, such as logistic regression. Nonetheless, regardless of the functional, the BLBB and SDBB approximate the summaries of the BB better as the size of the subsets $b = n^{\gamma}$ increases, so we encourage users to use subsets as large as possible. Another component that can affect the quality of the approximation is the number of bootstrap samples used to compute the summaries. We recommend setting this number as large as possible.

While the bias of the SDBB is smaller than the bias of the BLBB in approximating posterior means, the BLBB is better than the SDBB in approximating credible interval lengths and posterior standard deviations. The bias of the BLBB can be attributed to

Model	Summary	γ	BLBB	SDBB	ANS	AN
Linear	CIL	.6	.043	.088	.375	.046
		.7	.045	.062	.140	.046
		.8	.053	.048	.073	.046
	SD	.6	.054	.070	.368	.038
		.7	.041	.047	.134	.038
		.8	.034	.035	.066	.038
	Mean	.6	.003	.001	.003	<.001
		.7	.001	.001	.001	<.001
		.8	.001	.001	.001	<.001
Mixed	CIL	.6	.089	.107	.120	.063
		.7	.069	.093	.104	.063
		.8	.064	.056	.091	.063
	SD	.6	.086	.088	.123	.064
		.7	.064	.078	.107	.064
		.8	.049	.051	.093	.064
	Mean	.6	.002	.001	.002	<.001
		.7	.002	.001	.002	<.001
		.8	.001	.001	.002	<.001
Logistic	CIL	.6	.130	.208	.254	.022
		.7	.037	.093	.102	.022
		.8	.033	.050	.050	.022
	SD	.6	.172	.196	.252	.020
		.7	.075	.087	.101	.020
		.8	.047	.047	.048	.020
	Mean	.6	.470	.261	.464	.034
		.7	.235	.172	.244	.034
		.8	.175	.113	.164	.034

Table 1: Average relative and absolute errors of approximate posterior summaries for the functionals in (4)-(6). The average errors are computed over 100 simulated datasets and $(p+1)$ regression coefficients. Relative errors are reported for CIL and SD; absolute errors are reported for Mean. CIL and SD stand for credible interval length and posterior standard deviation.

the fact that it only considers a single partition of the dataset, which can be avoided by averaging over several random subsets (so the SDBB avoids this issue). However, the use of several random subsets adds an additional source of randomness that leads to wider credible intervals and larger standard deviations. Fortunately, as pointed out above, these differences are less worrisome as the size of the subsets increases.

3.2 Computational considerations

We compare the relative error of our methods with respect to 1,000 posterior draws from the BB (run on the full dataset), and we average our results over 100 simulated datasets.

In the case of the BLBB, the number of bootstrap samples within each subgroup equals 100, and the algorithm is run for 20 seconds (if all the subgroups are processed before 20 seconds, the algorithm stops). On the other hand, the SDBB runs for 20 seconds (unless 1,000 samples are sampled before 20 seconds, in which case the algorithm stops). For both methods, the subsets are processed sequentially using a single core. Figure 2 displays the results found for linear and logistic regression and credible interval lengths using $\gamma = 0.8$. Due to the similarity of conclusions, we do not present the results for mixed-effects regression and other values of γ.

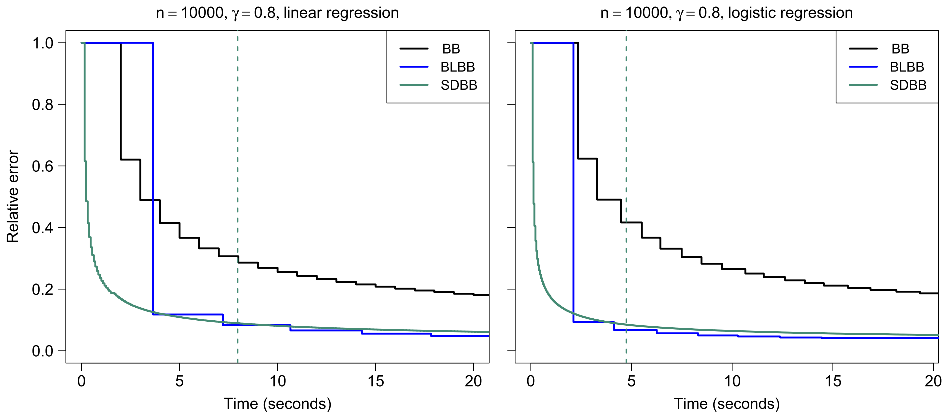

Figure 2: Average relative errors and processing times associated with the length of credible intervals for $\boldsymbol{\beta}^{\mathrm{BB}}_{\mathrm{lm},n}$ and $\boldsymbol{\beta}^{\mathrm{BB}}_{\mathrm{lg},n}$. The vertical green dashed line indicates when the SDBB has generated the 100th draw. The black line represents how fast the BB approximates the interval lengths based on 1,000 draws.

Figure 2 shows that the SDBB provides faster outputs than the BLBB if we do not wait until having 100 simulations to compute summaries of the posterior distribution. If we wait, the BLBB is faster than the SDBB. This phenomenon occurs because each iteration of the SDBB requires computing the functional twice, whereas the BLBB only computes the functional once. If we wait, the SDBB produces outputs just after the BLBB has processed the second subset. We recommend waiting until having some minimum number of samples because, otherwise, the estimates might not be reliable (in the sense that they will have high variance). Indeed, the figure shows that while the SDBB can produce outputs very quickly, reporting inferences with very few samples can have high relative error.

The computational cost of loading subsets of a massive dataset into memory is not always negligible and can play a crucial role in the performance of the proposed methods, particularly for the SDBB. If b is large enough, each subset takes a considerable amount of time to be loaded into memory, which can make the SDBB slower than the BLBB. More precisely, if we consider the same settings used in the simulation studies, the BLBB only has to load n/b subsets, while the SDBB needs to load 1,000 subsets into memory. Hence, the use of the SDBB is limited to scenarios where, according to the

Size of subgroup	Batch	Processing	Combining	Full BB
$n^{0.6}$	10	6	0.02	569
	100	96	0.46	966
$n^{0.7}$	10	22	0.02	2203
	100	152	0.22	1526
$n^{0.8}$	10	38	0.01	3865
	100	385	0.18	3854

Table 2: Average computation times (in seconds) for lossless inference for $\boldsymbol{\beta}^{\text{BB}}_{\text{lm},n}$ and using different subgroup and batch sizes.

available computational resources, the value of b is not too large or the data transfer rate is not too low.

Although the BLBB and SDBB do not start reporting results at the same time (which depends on the size of b and whether or not we impose a condition of having at least 100 draws from the SDBB), both provide results that stabilize rather quickly: in Figure 2, the relative errors tend to stabilize before the n/b subsets for the BLBB and the 1,000 subsets for the SDBB have been processed. This is appealing in scenarios with limited access to computational resources. Even if the user can load the full dataset into memory, our methods will stabilize faster and might output better results than the BB. ANS is a valid alternative in this context, but users have to be willing to accept higher relative errors.

We conclude this section discussing some computational aspects related to the lossless procedure presented in Section 2.3. While it is possible to perform lossless inference for $\boldsymbol{\beta}^{\text{BB}}_{\text{lm},n}$, it is significantly slower than the BLBB and SDBB. In our simulations, the BLBB and SDBB can produce excellent approximations in 20 seconds. Table 2 shows computing times for the lossless method with different subgroup sizes and number of bootstrap samples that are drawn at a time from the subgroups (referred to as "batch" in the table). We show the average time spent sampling from the subsets (processing) and the time it takes to combine the results once all the samples are drawn (combining). The last column is an estimate of how long the procedure would take to generate 1,000 samples using the lossless method. We observe that smaller subgroups and bigger batches are preferable, but in any case the time it takes to generate 1,000 samples is significantly larger than the time that it takes to produce them using the BLBB and SDBB (and recall that, in this context, the BLBB and SDBB provide very good approximations, so the benefit of using a lossless method is minimal).

4 Applications with real-life datasets

In this section, we apply the BLBB and SDBB to 2 different datasets. The first one is the Office of Personnel Management's Central Personnel Data File (https://www.opm.gov/), which we refer to as the OPM dataset from now on. The OPM dataset is confidential and housed by the Protected Research Data Network at Duke University https://oit.duke.edu/what-we-do/services/protected-network. We consider two subsets of the data: i) a subset comprising employees that worked during 2011 and is referred to as OPM-2011, and ii) a subset comprising employees that worked without any interruption during ten years starting in 2002, which we refer to as OPM-10Y. The second dataset includes public microdata files from the American Community Survey 2012 (ACS-2012), which can be downloaded from the United States Bureau of the Census (http://www2.census.gov/acs2012_1yr/pums/).

We study the performance of the BLBB and SDBB in approximating functionals of the posterior distribution of regression coefficients. We compare their 95% credible intervals, posterior standard deviations, and posterior means (as we did in Section 3). We approximate 95% credible intervals using $\xi_{l,4}^{\text{BLBB}} \pm (\xi_{l,2}^{\text{BLBB}} - \xi_{l,1}^{\text{BLBB}})/2$ and $\xi_{l,4}^{\text{SDBB}} \pm (\xi_{l,2}^{\text{SDBB}} - \xi_{l,1}^{\text{SDBB}})/2$.

We also compare the BLBB and SDBB to the asymptotic methods ANS and AN, which were defined in Section 3. For the OPM-2011 dataset, we focus on estimating coefficients from linear regression, i.e., $\boldsymbol{\beta}_{\text{lm},n}^{\text{BB}}$ (an analogous comparison for the τ-quantile regression estimator proposed in Chamberlain and Imbens (2003) can be found in the supplementary material). For the OPM-10Y dataset, we use mixed-effects regression and aim to approximate summaries of the distribution of $\boldsymbol{\beta}_{\text{mx},n}^{\text{BB}}$. For the ACS-2012 dataset, we estimate coefficients of a logistic regression model, which we denote $\boldsymbol{\beta}_{\text{lg},n}^{\text{BB}}$.

4.1 Office of Personnel Management

The OPM dataset comprises about 3.5 million employees and 29 variables recorded over 24 years. This dataset contains personnel records from employees that served in the federal U. S. government, including demographic variables (e.g., race, gender, and age) and relevant information related to their wages and careers (such as educational level). Our response variable is the natural logarithm of the wages, and the predictors correspond to the variables whose effect is of interest (e.g., gender or race) along with other variables that are used to control for potential confounding factors (e.g., age and education level). In order to assess wage gaps between the levels of a feature of interest (e.g., between men and women or between races), researchers have run regression models and interpreted their coefficients (see Bolton and de Figueiredo, 2016b,a; Barrientos et al., 2018). When those inferences cannot rely on parametric assumptions, uncertainty about the coefficients can be measured using nonparametric methods such as the BB.

For illustrative purposes, we use 2 random samples of 50,000 and 200,000 full-time employees from the OPM-2011 dataset, and 2 random samples of size 10,000 and 40,000 from the OPM-10Y dataset. We include gender, race, educational level, and age as predictors. The levels for gender are male and female, whereas the levels for race are white, American indian/Alaskan native, Asian or pacific islander, black, and hispanic. Age and educational level are treated as numerical variables, and we include both linear and quadratic effects on the age of the employees. The datasets contain 22 different educational categories. For ease of interpretation, we collapse the categories into a single continuous measure of the years of education attained by an individual past high school. Race is treated as a categorical variable, and the baseline cannot be disclosed because the dataset is confidential. For the OPM-10Y, we add a predictor representing the year at which the observation was collected. The regression coefficient $\boldsymbol{\beta}_{\text{mx},n}^{\text{BB}}$ is computed using Expression (6) where $\hat{V}$ is computed using a mixed-effects model with a random intercept and slope for year, and assuming that each employee represents a level of the grouping factor.

We compare the results obtained with the BLBB, SDBB, ANS, and AN with those obtained after drawing 1,000 samples from the BB ran on the full dataset. The analysis was performed on a computer with a Intel(R) Xeon(R) CPU E5-2699 v8 @ 2.20GHz processor and 236 GB RAM (running R version 3.3.3). We consider $b = n^\gamma$ with $\gamma = 0.6$, 0.7, and 0.8.

OPM-2011

A table containing the relative and absolute errors associated with $\boldsymbol{\beta}^{\text{BB}}_{\text{lm},n}$ can be found in the supplementary material. The relative errors for interval lengths and standard deviations are rather small (less than 0.05) with all methods. The biases (absolute errors) are also small (less than 0.015). In practice, the BLBB and ANS can give biased results, depending on the partition. In almost all of the cases, the BLBB and SDBB produced errors that are smaller or equal than the errors of ANS and AN.

In this application, the BLBB and SDBB provide satisfactory results for the values of n and γ we considered, and the effect of increasing n and γ is not particularly noticeable. On the other hand, the asymptotic methods improve their performance when n and γ are increased. Figure 3 displays credible intervals for coefficients associated with race. All the methods provide satisfactory approximations.

OPM-10Y

Table 3 contains the relative and absolute errors associated with $\boldsymbol{\beta}^{\text{BB}}_{\text{mx},n}$. The errors are rather large for $n = 10{,}000$ and $\gamma = 0.6$. However, as n or γ increase, the BLBB and SDBB are better approximations of the BB. For $n = 40{,}000$ and $\gamma = 0.8$, the errors are fairly small. For this specific dataset, none of the methods outperforms the others, but we observe some patterns. For example, for $\gamma = 0.6$, the BLBB assesses uncertainty better than the SDBB, and for $n = 10{,}000$, the bias associated with SDBB is smaller than the bias of the BLBB.

The relative errors for interval lengths and standard deviations are quite large with the asymptotic methods if we use the asymptotic estimator of the variance proposed in Jacqmin-Gadda et al. (2007). We also consider the sandwich estimator proposed in Liang and Zeger (1986), which is implemented in the R function `vcovCR.lme` of the library `clubSandwich` (Pustejovsky, 2018). The results with the sandwich estimator are labeled ANS-Sand and AN-Sand. Table 3 also contains the relative and absolute errors associated with these asymptotic methods. The errors associated with ANS-Sand are similar to the errors found with the BLBB.

Figure 3 displays credible intervals for coefficients associated with different levels of race. Again, we observe that increasing n or γ improves the quality of the approximation. For $n = 40{,}000$, we do not observe big discrepancies between the credible intervals based on the BB and the BLBB, SDBB, or ANS-Sand; however, the intervals with ANS tend to

n	Summary	γ	BLBB	SDBB	ANS	AN	ANS-Sand	AN-Sand
10,000	CIL	.6	.190	.307	.365	.259	.194	.009
		.7	.107	.128	.216	.259	.100	.009
		.8	.095	.044	.240	.259	.088	.009
	SD	.6	.197	.249	.370	.259	.198	.011
		.7	.100	.118	.218	.259	.100	.011
		.8	.087	.033	.240	.259	.084	.011
	Mean	.6	.204	.162	.204	.002	.204	.002
		.7	.055	.053	.054	.002	.054	.002
		.8	.024	.012	.023	.002	.023	.002
40,000	CIL	.6	.077	.105	.197	.269	.075	.009
		.7	.076	.024	.242	.269	.072	.009
		.8	.022	.027	.259	.269	.019	.009
	SD	.6	.067	.079	.199	.267	.068	.007
		.7	.067	.036	.239	.267	.064	.007
		.8	.010	.014	.257	.267	.011	.007
	Mean	.6	.055	.062	.054	<0.001	.054	<0.001
		.7	.008	.013	.008	<0.001	.008	<0.001
		.8	.011	.004	.011	<0.001	.011	<0.001

Table 3: Average relative and absolute errors of approximate posterior summaries for $\boldsymbol{\beta}^{\text{BB}}_{\text{mx},n}$, OMP-10Y dataset. The average errors are computed over all regression coefficients. Relative errors are reported for CIL and SD; absolute errors are reported for Mean. CIL and SD stand for credible interval length and posterior standard deviation.

be too narrow. Additionally, in Figure 3 we observe that the most problematic coefficient is the first one (represented as β_1 in the figure), which is associated with a category with a small observed frequency (2%). If a variable has some levels that are highly infrequent, approximating their coefficients via subsetting can be problematic: for example, some of the subsets might have extremely small frequencies for some levels, and some levels might not be observed at all.

In this specific context (OPM dataset, linear and mixed-effect regressions), our general recommendation to take subset sizes as large as possible leads to reasonable performance of the BLBB and the SDBB.

4.2 American Community Survey

The ACS-2012 dataset contains records of 1,477,091 households in the United States collected in 2012. The data were collected with the goal of making inferences about population demographics and housing characteristics. In this application, we only use the information at the household level and not at the individual level, and we model a binary outcome (response variable) that indicates whether or not the household is paying for a fire/hazard/flood insurance. We regress our binary outcome on a set of numerical and categorical predictors. As numerical predictors, we consider number of people living in the household, number of bedrooms, number of rooms, number of ve-

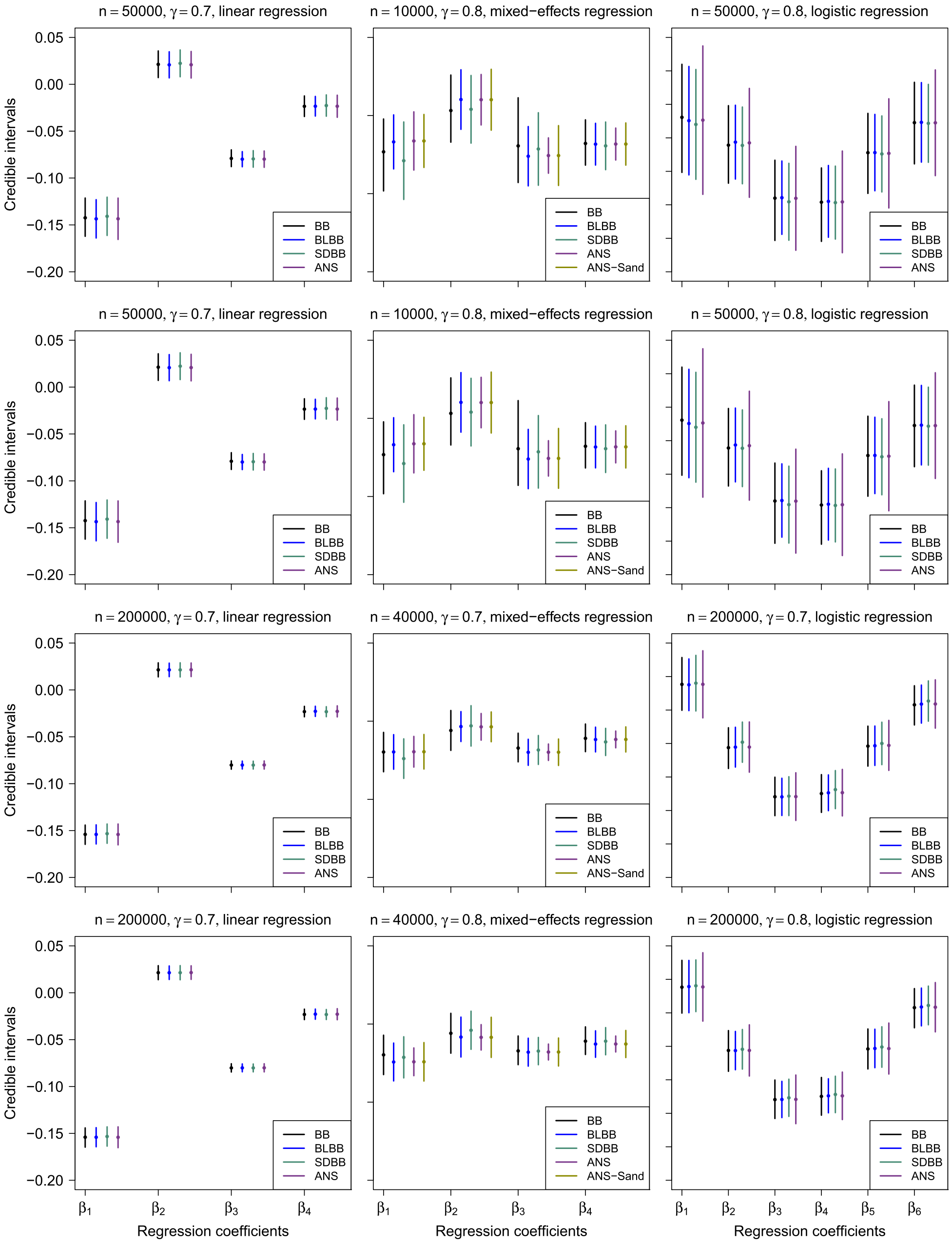

Figure 3: Credible intervals for regression coefficients, two different values of n, and $\gamma \in \{0.7, 0.8\}$. The first and second columns display intervals associated with the levels of the predictor race from the OPM-2011 and OPM-10Y datasets, respectively. The third column displays intervals associated with the levels of the predictor indicating how long the families lived in the household, ACS-2012 dataset.

hicles, and household income (in the past 12 months). The categorical predictors are lot size, yearly food stamp/Supplemental Nutrition Assistance Program (SNAP) recipiency, house heating fuel, presence and age of children, and a discretized variable that indicates how long the families lived in the household. This set of predictors leads to a regression with 21 coefficients.

We use two subsets of the ACS-2012 dataset of n = 50,000 and 200,000 complete cases records that correspond to households that are located in the Northeast of the United States and are not rentals. We estimate the posterior mean and standard deviation, quantiles (2.5% and 97.5%), and lengths of the resulting 95% credible intervals for $\boldsymbol{\beta}^{\mathrm{BB}}_{\mathrm{lg},n}$. We compare the results obtained with the BLBB, SDBB, and the asymptotic methods ANS and AN with those obtained from running 1,000 samples from the BB ran on the full dataset.

Table 4 contains the relative and absolute errors associated with $\boldsymbol{\beta}^{\mathrm{BB}}_{\mathrm{lg},n}$. The errors are moderately small (<0.09) even for n = 10,000 and $\gamma = 0.6$. We also observe that, as either n or γ increase, the BLBB and SDBB provide better results. In this application, we cannot conclude that the BLBB or SDBB uniformly outperforms the other; for $\gamma = 0.6$, the BLBB is better at assessing uncertainty than the SDBB, whereas for n = 50,000, the bias associated with SDBB is smaller than the bias of the BLBB. These patterns were also observed in the OPM-10Y dataset, where we used a mixed-effects model (see

n	Summary	γ	BLBB	SDBB	ANS	AN
50,000	CIL	.6	.033	.087	.844	.378
		.7	.034	.027	.286	.378
		.8	.045	.029	.301	.378
	SD	.6	.056	.082	.850	.374
		.7	.021	.026	.282	.374
		.8	.026	.024	.297	.374
	Mean	.6	.491	.414	.480	.010
		.7	.191	.168	.173	.010
		.8	.080	.069	.060	.010
200,000	CIL	.6	.020	.045	.307	.280
		.7	.034	.033	.218	.280
		.8	.036	.032	.240	.280
	SD	.6	.029	.041	.298	.271
		.7	.018	.023	.210	.271
		.8	.022	.027	.232	.271
	Mean	.6	.089	.118	.087	.009
		.7	.024	.063	.024	.009
		.8	.016	.021	.015	.009

Table 4: Average relative and absolute errors of approximate posterior summaries for $\boldsymbol{\beta}^{\mathrm{BB}}_{\mathrm{lg},n}$, ACS-2012 dataset. The average errors are computed over all regression coefficients. Relative errors are reported for CIL and SD; absolute errors are reported for Mean. CIL and SD stand for credible interval length and posterior standard deviation.

Section 4.1). The relative errors for estimating interval lengths and standard deviations with the ANS and AN are large. The errors decrease as n increases, but they are never smaller than the errors of the BLBB and SDBB. The bias (absolute error) associated with AN is very small, whereas the bias associated with ANS is similar to the bias of the BLBB. This is not surprising because we are using the same partition with both methods (BLBB and ANS).

Figure 3 displays credible intervals for the coefficients of the variable that indicates how long the families lived in the household. We choose to show these intervals because the observed frequencies of some of the levels are low, so they are particularly hard to estimate with subsetting methods. ANS is more sensitive to this specific issue (which is even worse when b is small) than the BLBB and SDBB. In general, we observe that ANS tends to output intervals that are too wide.

5 Discussion

We have presented the BLBB and SDBB as two data-subsetting procedures to approximate the BB. The BLBB and SDBB are analogous to the BLB (Kleiner et al., 2014) and SDB (Sengupta et al., 2016). The proposed procedures have theoretical and computational properties that are comparable to those of their frequentist counterparts. The performance of the methods has been illustrated and compared in simulation studies and real datasets. Although both the BLBB and SDBB are computationally efficient, the BLBB is preferable in scenarios where the computational cost of loading the subsets into memory is high. A similar conclusion can be drawn if the BLB and SDB are compared in an analogous setting. We observe that the BLBB approximates the uncertainty of the BB better than the SDBB, whereas the SDBB provides better approximations of point estimates than the BLBB. If the subsets can be loaded into memory reasonably fast and the functional of interest can be computed quickly, we recommend running rescaled bootstraps for some subsets to check if the posterior distributions of the functionals are similar. If they are, the SDBB will approximate the uncertainty as well as the BLBB does but with less bias; if they are not, the SDBB will overestimate uncertainty estimates and we recommend using the BLBB.

The performance of the methods depends on the size of the subsets and the functional of interest. In general, we observe that increasing subset sizes improves the approximation. This relationship between the quality of the approximation and subset size is not particular to our procedures; in fact, it is a common issue of data-subsetting methods. In addition to the BLBB and SDBB, we provide a strategy for performing lossless inference for functionals that can be expressed as functions of expectations with respect to the probability measure of the BB. This class is larger than one would expect at first glance: it includes, for instance, the weighted least squares estimator used in Clyde and Lee (2001) and Taddy et al. (2016), as well as the instrumental variables estimator introduced in Section 2 of Chamberlain and Imbens (2003).

Future work can extend our contribution. For example, it would be useful to determine which functionals are best estimated by the BLBB or SDBB (beyond empirical investigations), so that we can select and combine the methods as needed depending on the functionals we want to estimate. It would also be interesting to find strategies for determining when the sample size is big enough so that asymptotic methods (such as ANS, as defined in Section 3) can be used to our advantage. Another interesting direction for further research would be designing data-subsetting strategies for datasets that have categorical variables with low observed frequency levels, which is an important practical issue, as we argue in Section 4.1.

Supplementary Material

Supplementary material: Bayesian Bootstraps for Massive Data (DOI: 10.1214/19-BA1155SUPP; .pdf). The supplementary material has 6 sections: the first provides theoretical results for the processes proposed in Sections 2.1, 2.2, and 2.3; the second has a figure which details the Monte Carlo algorithm for performing lossless inference for the class of functionals described in Section 2.3; the third contains a scheme for lossless simulation for the example in Section 2.3 in Chamberlain and Imbens (2003); the fourth part explains how to perform lossless inference for the Dirichlet-Multinomial process; the fifth includes a table with relative and absolute errors related to the linear regression coefficients estimated from the OPM-2011 dataset in Section 4.1; Finally, the sixth assesses the performance of the BLBB, SDBB, ANS, and AN approximating coefficients of a quantile regression fitted to the OPM-2011 dataset.

References

Barrientos, A. F. and Peña, V. (2019). "Supplementary material: Bayesian bootstraps for massive data." *Bayesian Analysis.* doi: https://doi.org/10.1214/19-BA1155SUPP. 367

Barrientos, A. F., Bolton, A., Balmat, T., Reiter, J. P., de Figueiredo, J. M., Machanavajjhala, A., Chen, Y., Kneifel, C., and DeLong, M. (2018). "Providing access to confidential research data through synthesis and verification: An application to data on employees of the U.S. federal government." *The Annals of Applied Statistics*, 12(2): 1124–1156. MR3834297. doi: https://doi.org/10.1214/18-AOAS1194. 377

Bolton, A. D. and de Figueiredo, J. M. (2016a). "Measuring and explaining the gender wage Gap in the federal government." Paper presented at the 2016 annual meeting of the American Political Science Association, Philadelphia, Pennsylvania. 377

Bolton, A. D. and de Figueiredo, J. M. (2016b). "Rising wages and human capital in the federal government." Paper presented at the 2016 annual meeting of the Southern Political Science Association, San Juan, Puerto Rico. 377

Carroll, R. J. and Pederson, S. (1993). "On robustness in the logistic regression model." *Journal of the Royal Statistical Society. Series B (Methodological)*, 693–706. MR1223937. 371

Castillo, I. and Nickl, R. (2014). "On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures." *The Annals of Statistics*, 42(5): 1941–1969. MR3262473. doi: https://doi.org/10.1214/14-AOS1246. 370

Chamberlain, G. and Imbens, G. W. (2003). "Nonparametric applications of Bayesian inference." *Journal of Business & Economic Statistics*, 21(1): 12–18. MR1973803. doi: https://doi.org/10.1198/073500102288618711. 369, 377, 382, 383

Choudhuri, N. (1998). "Bayesian bootstrap credible sets for multidimensional mean functional." *The Annals of Statistics*, 26(6): 2104–2127. MR1700223. doi: https://doi.org/10.1214/aos/1024691463. 365

Cifarelli, D. M. and Melilli, E. (2000). "Some new results for Dirichlet priors." *The Annals of Statistics*, 28(5): 1390–1413. MR1805789. doi: https://doi.org/10.1214/aos/1015957399. 365

Clyde, M. and Lee, H. (2001). "Bagging and the Bayesian bootstrap." In Richardson, T. and Jaakkola, T. (eds.), *Artificial Intelligence and Statistics*, 169–174. 365, 369, 371, 382

Datta, J. and Ghosh, J. K. (2014). "Bootstrap—an exploration." *Statistical Methodology*, 20: 63–72. MR3205721. doi: https://doi.org/10.1016/j.stamet.2013.08.003. 365

Dong, Q., Elliott, M. R., and Raghunathan, T. E. (2014). "A nonparametric method to generate synthetic populations to adjust for complex sampling design features." *Survey Methodology*, 40(1): 29–46. 365

Efron, B. (1979). "Bootstrap methods: another look at the jackknife." *The Annals of Statistics*, 7(1): 1–26. MR0515681. 364

Fushiki, T. (2010). "Bayesian bootstrap prediction." *Journal of Statistical Planning and Inference*, 140(1): 65–74. MR2568122. doi: https://doi.org/10.1016/j.jspi.2009.06.007. 365

Gasparini, M. (1995). "Exact multivariate Bayesian bootstrap distributions of moments." *The Annals of Statistics*, 23(3): 762–768. MR1345198. doi: https://doi.org/10.1214/aos/1176324620. 365

Graham, D. J., McCoy, E. J., and Stephens, D. A. (2016). "Approximate Bayesian inference for doubly robust estimation." *Bayesian Analysis*, 11(1): 47–69. MR3447091. doi: https://doi.org/10.1214/14-BA928. 365

Gu, J., Ghosal, S., and Roy, A. (2008). "Bayesian bootstrap estimation of ROC curve." *Statistics in Medicine*, 27(26): 5407–5420. MR2542360. doi: https://doi.org/10.1002/sim.3366. 365

Hahn, J. (1997). "Bayesian bootstrap of the quantile regression estimator: a large sample study." *International Economic Review*, 38(4): 795–808. MR1482116. doi: https://doi.org/10.2307/2527216. 365

Heckelei, T. and Mittelhammer, R. C. (2003). "Bayesian bootstrap multivariate regression." *Journal of Econometrics*, 112(2): 241–264. MR1951144. doi: https://doi.org/10.1016/S0304-4076(02)00196-3. 365

Ishwaran, H., James, L. F., and Zarepour, M. (2009). "An alternative to the m out of n bootstrap." *Journal of statistical planning and inference*, 139(3): 788–801. MR2479828. doi: https://doi.org/10.1016/j.jspi.2008.04.032. 365

Ishwaran, H. and Zarepour, M. (2002). "Exact and approximate sum representations for the Dirichlet process." *Canadian Journal of Statistics*, 30(2): 269–283. MR1926065. doi: https://doi.org/10.2307/3315951. 370

Jacqmin-Gadda, H., Sibillot, S., Proust, C., Molina, J.-M., and Thiébaut, R. (2007). "Robustness of the linear mixed model to misspecified error distribution." *Computational Statistics & Data Analysis*, 51(10): 5142–5154. MR2370713. doi: https://doi.org/10.1016/j.csda.2006.05.021. 371, 378

James, L. F. (1997). "A study of a class of weighted bootstraps for censored data." *The Annals of Statistics*, 25(4): 1595–1621. MR1463566. doi: https://doi.org/10.1214/aos/1031594733. 365

James, L. F. (2008). "Large sample asymptotics for the two-parameter Poisson–Dirichlet process." In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, 187–199. Institute of Mathematical Statistics. 370

Kim, Y. and Lee, J. (2003). "Bayesian bootstrap for proportional hazards models." *The Annals of Statistics*, 31(6): 1905–1922. MR2036394. doi: https://doi.org/10.1214/aos/1074290331. 365

Kingman, J. F. (1975). "Random discrete distributions." *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–22. MR0368264. 370

Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M. I. (2014). "A scalable bootstrap for massive data." *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 76(4): 795–816. MR3248677. doi: https://doi.org/10.1111/rssb.12050. 363, 366, 367, 382

Lee, H. K. H. and Clyde, M. A. (2004). "Lossless online Bayesian bagging." *Journal of Machine Learning Research*, 5: 143–151. MR2247976. 369

Li, C., Srivastava, S., and Dunson, D. B. (2017). "Simple, scalable and accurate posterior interval estimation." *Biometrika*, 104(3): 665–680. MR3694589. doi: https://doi.org/10.1093/biomet/asx033. 364

Liang, K.-Y. and Zeger, S. L. (1986). "Longitudinal data analysis using generalized linear models." *Biometrika*, 73(1): 13–22. MR0836430. doi: https://doi.org/10.1093/biomet/73.1.13. 378

Lo, A. Y. (1983). "Weak convergence for Dirichlet processes." *Sankhyā: The Indian Journal of Statistics, Series A*, 105–111. MR0749358. 370

Lo, A. Y. (1987). "A large sample study of the Bayesian bootstrap." *The Annals of Statistics*, 15(1): 360–375. MR0885742. doi: https://doi.org/10.1214/aos/1176350271. 365

Lo, A. Y. (1988). "A Bayesian bootstrap for a finite population." *The Annals of Statistics*, 16(4): 1684–1695. MR0964946. doi: https://doi.org/10.1214/aos/1176351061. 365

Lo, A. Y. (1991). “Bayesian bootstrap clones and a biometry function.” *Sankhyā Ser. A*, 53(3): 320–333. MR1189775. 365, 366

Lo, A. Y. (1993). “A Bayesian bootstrap for censored data.” *The Annals of Statistics*, 21(1): 100–123. MR1212168. doi: https://doi.org/10.1214/aos/1176349017. 365

Lyddon, S., Holmes, C., and Walker, S. (2019). “Generalized Bayesian updating and the loss-likelihood bootstrap.” *Biometrika* (in press). 364

Meeden, G. (1993). “Noninformative nonparametric Bayesian estimation of quantiles.” *Statistics & Probability Letters*, 16(2): 103–109. MR1212042. doi: https://doi.org/10.1016/0167-7152(93)90153-A. 365

Minsker, S., Srivastava, S., Lin, L., and Dunson, D. B. (2017). “Robust and scalable Bayes via a median of subset posterior measures.” *The Journal of Machine Learning Research*, 18(1): 4488–4527. MR3763758. 364, 366

Muliere, P. and Secchi, P. (1996). “Bayesian nonparametric predictive inference and bootstrap techniques.” *Annals of the Institute of Statistical Mathematics*, 48(4): 663–673. MR1440263. doi: https://doi.org/10.1007/BF00052326. 365, 370

Neiswanger, W., Wang, C., and Xing, E. P. (2014). “Asymptotically exact, embarrassingly parallel MCMC.” In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI’14, 623–632. AUAI Press. 364

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2018). *nlme: Linear and nonlinear mixed effects models*. R package version 3.1-137. URL https://CRAN.R-project.org/package=nlme 373

Pitman, J. (1995). “Exchangeable and partially exchangeable random partitions.” *Probability Theory and Related Fields*, 102: 145–158. MR1337249. doi: https://doi.org/10.1007/BF01213386. 370

Pitman, J. (1996). “Some developments of the Blackwell-MacQueen urn scheme.” *Lecture Notes-Monograph Series*, 245–267. MR1481784. doi: https://doi.org/10.1214/lnms/1215453576. 370

Pustejovsky, J. (2018). *clubSandwich: Cluster-robust (Sandwich) variance estimators with small-sample corrections*. R package version 0.3.2. URL https://CRAN.R-project.org/package=clubSandwich 378

R Core Team (2015). “R: A language and environment for statistical computing.” URL https://www.R-project.org/ 373

Rubin, D. B. (1981). “The Bayesian bootstrap.” *The Annals of Statistics*, 9(1): 130–134. MR0600538. 364

Rubin, D. B. and Schenker, N. (1986). “Multiple imputation for interval estimation from simple random samples with ignorable nonresponse.” *Journal of the American Statistical Association*, 81(394): 366–374. MR0845877. 365

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I., and Mc-

Culloch, R. E. (2016). "Bayes and big data: The consensus Monte Carlo algorithm." *International Journal of Management Science and Engineering Management*, 11(2): 78–88. 364

Sengupta, S., Volgushev, S., and Shao, X. (2016). "A subsampled double Bootstrap for massive data." *Journal of the American Statistical Association*, 111(515): 1222–1232. MR3561944. doi: `https://doi.org/10.1080/01621459.2015.1080709`. 363, 367, 368, 382

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Statistica Sinica*, 2: 639–650. MR1309433. 370

Siddique, J. and Belin, T. R. (2008). "Using an approximate Bayesian bootstrap to multiply impute nonignorable missing data." *Computational Statistics & Data Analysis*, 53(2): 405–415. MR2649095. doi: `https://doi.org/10.1016/j.csda.2008.07.042`. 365

Srivastava, S., Cevher, V., Tran-Dinh, Q., and Dunson, D. B. (2015). "WASP: Scalable Bayes via barycenters of subset posteriors." In *Artificial Intelligence and Statistics*. 364, 366

Srivastava, S., Li, C., and Dunson, D. B. (2018). "Scalable Bayes via barycenter in Wasserstein space." *The Journal of Machine Learning Research*, 19(1): 312–346. MR3862415. 364

Taddy, M., Chen, C.-S., Yu, J., and Wyle, M. (2015). "Bayesian and empirical Bayesian forests." In Blei, D. and Bach, F. (eds.), *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, 967–976. 364, 365

Taddy, M., Gardner, M., Chen, L., and Draper, D. (2016). "A nonparametric Bayesian analysis of heterogeneous treatment effects in digital experimentation." *Journal of Business & Economic Statistics*, 34(4): 661–672. MR3548002. doi: `https://doi.org/10.1080/07350015.2016.1172013`. 364, 365, 369, 371, 382

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics. MR1385671. doi: `https://doi.org/10.1007/978-1-4757-2545-2`. 367

Varron, D. (2014). "Donsker and Glivenko-Cantelli theorems for a class of processes generalizing the empirical process." *Electronic Journal of Statistics*, 8(2): 2296–2320. MR3275745. doi: `https://doi.org/10.1214/14-EJS955`. 370

Wang, X. and Dunson, D. B. (2013). "Parallelizing MCMC via Weierstrass sampler." *arXiv preprint* arXiv:1312.4605. 364

Wang, X., Guo, F., Heller, K. A., and Dunson, D. B. (2015). "Parallelizing MCMC with random partition trees." In *Advances in Neural Information Processing Systems*, 451–459. 364

Welsh, A. and Richardson, A. (1997). "13 Approaches to the robust estimation of mixed models." *Handbook of Statistics*, 15: 343–384. MR1492718. doi: `https://doi.org/10.1016/S0169-7161(97)15015-5`. 371

Weng, C.-S. (1989). "On a second-order asymptotic property of the Bayesian bootstrap mean." *The Annals of Statistics*, 17(2): 705–710. MR0994261. doi: https://doi.org/10.1214/aos/1176347136. 365

Zhou, H., Elliott, M. R., and Raghunathan, T. E. (2016). "Multiple imputation in two-stage cluster samples using the weighted finite population Bayesian bootstrap." *Journal of Survey Statistics and Methodology*, 4(2): 139–170. 365

Acknowledgments

We would like to thank Jerome P. Reiter, Lancelot F. James, and Peter Müller for helpful comments that improved the content and presentation of the article. We would also like to thank Alexander D. Bolton and John M. de Figueiredo for providing the OPM dataset and their valuable feedback. A. F. Barrientos was supported by grants from the National Science Foundation (ACI 1443014 and SES 1131897) and the Alfred P. Sloan Foundation (G-2-15-20166003).