Finite Sample Valid Inference via Calibrated Bootstrap

Yiran Jiang
Department of Biostatistics, Yale University, New Haven, USA
Chuanhai Liu†
Department of Statistics, Purdue University, West Lafayette, USA
Heping Zhang
Department of Biostatistics, Yale University, New Haven, USA

Summary. While widely used as a general method for uncertainty quantification, the bootstrap method encounters difficulties that raise concerns about its validity in practical applications. This paper introduces a new resampling-based method, termed *calibrated bootstrap*, designed to generate finite sample-valid parametric inference from a sample of size n. The central idea is to calibrate an m-out-of-n resampling scheme, where the calibration parameter m is determined against inferential pivotal quantities derived from the cumulative distribution functions of loss functions in parameter estimation. The method comprises two algorithms. The first, named *resampling approximation* (RA), employs a *stochastic approximation* algorithm to find the value of the calibration parameter $m = m_\alpha$ for a given α in a manner that ensures the resulting m-out-of-n bootstrapped $1-\alpha$ confidence set is valid. The second algorithm, termed *distributional resampling* (DR), is developed to further select samples of bootstrapped estimates from the RA step when constructing $1-\alpha$ confidence sets for a range of α values is of interest. The proposed method is illustrated and compared to existing methods using linear regression with and without L_1 penalty, within the context of a high-dimensional setting and a real-world data application. The paper concludes with remarks on a few open problems worthy of consideration.

Keywords: Bayesian Bootstrap; L_1 penalty; Pivotal quantities; Profile likelihood; Stochastic approximation.

1. Introduction

Bootstrap methods are designed to estimate the sample distribution of the statistic of interest by resampling the observed data. Since the seminal paper of Efron (1979) (see also Efron, 2003), these methods have often served as an indispensable tool in the realm of statistical inference and prediction, due to their simplicity and efficiency (see, *e.g.* Efron and Tibshirani, 1993). It has also motivated relevant research in other contexts, including its Bayesian counterpart (see Rubin, 1981; Efron, 1982; Newton and Raftery, 1994; Efron, 2012; Newton et al., 2021, and references therein).

Nevertheless, practitioners may misuse the bootstrap methods in situations where no theoretical confirmation has been made (Shao and Tu, 2012). Care must be taken

†Corresponding author. Email: chuanhai@purdue.edu

in specific applications (Martin, 2015). Firstly, the theoretical underpinnings of the bootstrap methods predominantly center around its asymptotic validity. Specifically, researchers primarily concentrate on establishing the property of consistent estimation for the sampling distribution of the statistic of interest as the sample size approaches infinity (see, *e.g.*, Bickel and Freedman, 1981; Singh, 1981; Wellner and Zhan, 1996; van der Vaart and Wellner, 1996; Kosorok, 2008; Efron and Tibshirani, 1993, and references therein). Secondly, it has been reported that bootstrap can fail, even in its intended applications based on large-sample theory (see, *e.g.,* Bickel and Freedman, 1983; Abrevaya and Huang, 2005; Chernozhukov et al., 2023, and references therein). For example, the bootstrap has been shown to provide an inconsistent estimation of the true sampling distribution of the coefficients in high-dimensional linear regression setting of $n \to \infty$ while $p/n \to c$ for some $c > 0$ with p predictors, and the numerical experiments also show that the existing bootstrap methods give very poor inference on the vector of regression coefficients β (El Karoui and Purdom, 2018).

To overcome the limitations of the standard bootstrap method (*i.e.*, the bootstrap with n-out-of-n replacement resampling) in certain scenarios, researchers have explored alternative resampling approaches. These alternatives, also required to be supported by their asymptotic validity, include m-out-of-n resampling with replacement and without replacement, where the latter is also known as subset sampling (Politis and Romano, 1994; Bickel and Sakov, 2008; Bickel et al., 2012), an idea that dates back to Bretagnolle (1983). However, to the best of our knowledge, there is still a significant gap in the existing research when it comes to the development of theoretically supported finite sample inference methods based on resampling.

In order to achieve desirable finite-sample performance, the resampling scheme may need to adapt itself according to the specific model and observations. In this paper, we explore adaptive and computationally feasible resampling methods for achieving valid frequentist inference for model parameters in the finite sample case. That is, for a given model, the inference method should be able to provide control over the type-I error rates. For example, the constructed 95% confidence interval of the parameter should cover the true parameter at least (and close to) 95% of the time. For this, we develop computational methods with sound theoretical justifications, along the lines of recent developments on the foundations of statistical inference (Fisher, 1973; Shafer, 1976; Dempster, 2008; Singh et al., 2007; Xie and Singh, 2013; Hannig, 2009; Hannig et al., 2016; Martin and Liu, 2013, 2015; Martin, 2015; Cella and Martin, 2022), while our exposition will make use of the basic idea of the familiar pivotal method as much as possible.

Theoretically sound methods for finite-sample valid inference have been well-studied in the existing literature. However, the computational tools that can be practically applied to find solutions are generally lacking and often infeasible, especially in high-dimensional scenarios (Martin, 2015). The primary focus of this paper is to propose a computationally efficient, resampling-based numerical strategy to closely approximate the targeted theoretical solution. Philosophically, it is at least subtle to make inference via resampling because variability from resampling and uncertainty assessment for inference are two different concepts. Here, our key idea is to match the variability from resampling to uncertainty in inference. By looking at inferential problems from the inferential models (IMs) perspective (Martin and Liu, 2013), which could be considered as rooted in the piv-

otal method for hypothesis testing and confidence interval construction, we numerically quantify uncertainty assessment by considering loss functions in the context of parameter estimation and making use of the cumulative distributions of their sampling distributions to introduce the needed pivotal quantities or auxiliary random variables. Consequently, we seek resampling schemes in such a way that the distribution of the loss function based on the bootstrapped estimates matches the theoretical distribution necessary for valid inference. Because our primary goal is to create a method that leads to valid frequentist inference in finite sample scenarios, regardless of the asymptotic relationship between the sample size n and the number of parameters p, the method inherently possesses the capacity to be applicable in high-dimensional settings, for which the standard bootstrap has been known to be problematic as discussed above.

The rest of the paper is arranged as follows. In Section 2, we will briefly review the generalized association method (Martin, 2015), a foundational approach for valid parametric inference, which forms the theoretical basis of our proposed method. In Section 3, we propose a variant of bootstrap that aims to produce valid frequentist inference in finite sample scenarios. This method, called *calibrated bootstrap* (CB), consists of two steps. The first step applies a resampling method, called the *resampling approximation* (RA), to a set of pre-specified coverage probabilities. For each confidence coefficient, RA performs the m-out-of-n resampling with replacement, with the value of m obtained by a *stochastic approximation* algorithm to ensure that the resulting confidence region has the desired coverage probability. The second step of CB, called the *distributional resampling* (DR), is optional and selects a common sample of the bootstrapped estimates that works for a pre-specified range of α values of interest. Although our exposition emphasizes joint inference on the unknown parameters, we also discuss marginal inference on a parameter of interest. The proposed CB method is illustrated with applications in both easy-to-verify linear regression in Section 3 and challenging L_1 penalized linear regression in Section 4, including a real data example from a diabetes study. We conclude with a few remarks in Section 5.

2. Foundations of Valid Inference: an Overview

2.1. The Generalized Association Method for Parametric Inference

Here, we provide a brief review of the valid inference discussed in Martin (2015), which will be taken as our starting point in Section 3, where our proposed method will be introduced. Let y be an observed sample of size n from the true population $Y \sim \mathbf{P}_\theta$, $\theta \in \Theta$. Suppose that we are interested in inference on θ. We consider the problem of estimating θ with some loss function $\ell(y, \theta)$. Here, for our exposition, we focus on the likelihood-based loss function

$$\ell(y, \theta) = -\ln L_y(\theta) + \pi(\theta), \qquad \theta \in \Theta \tag{1}$$

where $L_y(\theta)$ denotes the likelihood function and $\pi(\theta)$ stands for a penalty function. Following Martin (2015, see Equation (1)), we take the *generalized association function*

$$T_{y,\theta} = \ell(y, \hat{\theta}_y) - \ell(y, \theta), \qquad \theta \in \Theta, y \in \mathbb{Y} \tag{2}$$

where $\hat{\theta}_y = \arg\min_\theta \ell(y, \theta)$ and it is assumed that $\min_\theta \ell(y, \theta)$ is finite.

As discussed in Martin (2015) and Martin and Liu (2015), valid inference can be made by making use of the distribution of $T_{Y,\theta}$,

$$F_\theta(t) = \mathrm{P}_{Y|\theta}\left(T_{Y,\theta} \le t\right), \tag{3}$$

where $Y \sim \mathbf{P}_\theta$ and $\theta \in \Theta$ is the fixed parameter value. The basic idea behind inferential models is that based on the fact that given the value U of $F_\theta\left(T_{y,\theta}\right)$, the knowledge we know about θ from the observed data y is exactly represented by the set

$$\left\{\theta:\ F_\theta\left(T_{y,\theta}\right) = U\right\}.$$

This leads to an exact frequentist confidence region $\{\theta:\ F_\theta(T_{y,\theta}) \ge \alpha\}$ with coverage probability $1-\alpha$ for all $\alpha \in (0,1)$. More precisely, we have the following theoretical result (*c.f.*, Martin, 2015).

Theorem 1 (Exact confidence region). *If $T_{Y,\theta}$ is a continuous random variable as a function of $Y \sim \mathbf{P}_\theta$, for any $\alpha \in (0,1)$, the set $\{\theta:\ F_\theta(T_{Y,\theta}) \ge \alpha\}$ is an exact $1-\alpha$ frequentist confidence region. Namely, for $Y \sim \mathbf{P}_\theta$ and the fixed parameter value $\theta \in \Theta$,*

$$\mathrm{P}_{Y|\theta}\left(\theta \notin \{\theta:\ F_\theta(T_{Y,\theta}) \ge \alpha\}\right) = \alpha$$

Proof. If $T_{Y,\theta}$ is a continuous random variable as a function of $Y \sim \mathbf{P}_\theta$, the distribution of $F_\theta(T_{Y,\theta})$ follows the standard uniform distribution Uniform$(0,1)$. Therefore,

$$\begin{aligned}\mathrm{P}_{Y|\theta}\left(\theta \notin \{\theta:\ F_\theta(T_{Y,\theta}) \ge \alpha\}\right) &= \mathrm{P}_{Y|\theta}\left(F_\theta(T_{Y,\theta}) \le \alpha\right)\\ &= \alpha,\end{aligned}$$

completing the proof.

For the sake of clarity in the context of constructing confidence regions, we formally define the validity of a confidence-oriented inference procedure, which is consistent with that used in the bootstrap literature for considering large sample-based validity.

Definition 1 (Valid Parametric Inference). *A confidence-oriented inference procedure for the model parameter $\theta \in \Theta$ is said to be valid at level $\alpha \in (0,1)$ if its $1-\alpha$ confidence region $\mathbf{C}(Y)$ satisfies*

$$\mathrm{P}_{Y|\theta}\left(\theta \notin \mathbf{C}(Y)\right) \le \alpha$$

as $Y \sim \mathbf{P}_\theta$. It is said to be valid if it is valid at all levels.

Thus, from Definition 1 and Theorem 1, we see that the key to obtaining the valid inference based on $T_{y,\theta}$ is to be able to evaluate $F_\theta(t)$ defined in (3). Here is a simple example where $F_\theta(t)$ can be obtained analytically.

Remark 1. *The pioneering work of Martin (2015), developed in the framework of IMs, can be viewed as to use the set $\{\theta:\ F_\theta(T_{y,\theta}) \ge \alpha\}$ to construct an exact confidence region for θ. However, it is not always feasible to derive a closed-form expression of $F_\theta(T_{y,\theta})$, necessitating Monte Carlo estimation. Meanwhile, evaluating the function value across a wide range of θ is required. Consequently, as is pointed out by Martin (2015), the computational cost can become prohibitively high, especially when dealing with high-dimensional θ.*

REMARK 2. *Besides* (2)*, the generalized association function* $T_{y,\theta}$ *can take other forms (Martin, 2015). Although the algorithm proposed in Section 3 does not depend on this specific formulation, we choose to adhere to* (2) *throughout our discussion. The choice is motivated by the fact that (2) conveniently summarizes all information in* y *concerning* θ*, and (2) is computationally simple compared to other potential forms of* $T_{y,\theta}$*, such as the standardized signed log likelihood ratio (Barndorff-Nielsen, 1986).*

2.2. T-Confidence Distribution

Utilizing the valid confidence set $\{\theta : \ F_\theta(T_{y,\theta}) \geq \alpha\}$ discussed in Section 2.1, suppose now we are interested in constructing a distribution of a function of both the unknown parameter θ and the data y, where y is taken as random variable, such that it assigns at least probability $1-\alpha$ to all the $100(1-\alpha)\%$ confidence sets. Intuitively, for each fixed y, our goal is to create a distribution of the parameter θ, denoted by G_y, such that for samples drawn from G_y and sorted according to the value $F_\theta(T_{y,\theta})$, the proportion of these values greater than α is at least $1-\alpha$ for any $\alpha \in (0,1)$. By Theorem 1, this guarantees the desired property of the distribution G_y. Such a distribution will greatly facilitate the inference procedure, as the choice of α and the confidence set is arbitrary. Formally, we define such a distribution as a *T-confidence distribution* as follows (*c.f.*, Martin, 2021b, 2023a):

DEFINITION 2 (T-CONFIDENCE DISTRIBUTION). *Given the observed data* y *from* $Y \sim \mathbf{P}_\theta$ *and a function* $T_{y,\theta}$ *of* y *and* θ*, a distribution* G_y*, indexed by* y *and on the space of* θ*, is said to be a* T*-confidence distribution of the unknown parameter* θ *with respect to* $T_{y,\theta}$ *if* G_y *satisfies:*

$$\mathrm{P}_{\theta^*}\left(F_{\theta_*}(T_{y,\theta_*}) \leq \alpha\right) = \alpha, \quad \textit{for any } \alpha \in (0,1), \tag{4}$$

or equivalently, $F_{\theta_*}(T_{y,\theta_*}) \sim$ *Uniform*$(0,1)$*, when* $\theta_* \sim G_y$*.*

For simplicity, hereafter we refer to the T-confidence distribution defined above simply as the *confidence distribution*. It follows that confidence sets can be produced straightforwardly from confidence distributions, as shown in the following theorem.

THEOREM 2. *Suppose that* G_y *is a confidence distribution with respect to* $T_{y,\theta}$ *for given the observed data* y *from* $Y \sim \mathbf{P}_\theta$*. For any* $\alpha \in (0,1)$*, let* $U_y^{1-\alpha}$ *be the subset of* Θ *determined by* $\int_{U_y^{1-\alpha}} dG_y \geq 1-\alpha$*. Then* $U_y^{1-\alpha}$*, as a confidence set for the true parameter* θ*, has at least* $100(1-\alpha)\%$ *coverage probability, that is, for* $Y \sim \mathbf{P}_\theta$*,*

$$\mathrm{P}_{Y|\theta}\left(\theta \notin U_Y^{1-\alpha}\right) \leq \alpha.$$

PROOF. The function mapping $\theta \mapsto F_\theta(T_{y,\theta})$ gives $U_y^{1-\alpha} \mapsto S_y^{1-\alpha} \subset (0,1)$, with $\mathrm{P}_{\theta^*}\left(F_{\theta_*}(T_{y,\theta_*}) \in S_y^{1-\alpha}\right) \geq 1-\alpha$. Since $F_{\theta_*}(T_{y,\theta_*})$ follows the standard uniform distribution denoted by U, we have $\int_{S_y^{1-\alpha}} dU \geq 1-\alpha$. It follows from Theorem 1 that, when $Y \sim \mathbf{P}_\theta$,

$$\mathrm{P}_{Y|\theta}\left(\theta \notin \{\theta : \ F_\theta(T_{Y,\theta}) \in S_Y^{1-\alpha}\}\right) = \mathrm{P}_{Y|\theta}\left(F_\theta(T_{Y,\theta}) \notin S_Y^{1-\alpha}\right) \leq \alpha,$$

completing the proof.

REMARK 3. *A more rigorous definition of the confidence distribution from the imprecise probability point of view (Martin, 2021a, 2023a) is given in Supplementary S.1. Furthermore, Martin (2023a) shows that if well-defined, the fiducial distribution obtained via the traditional fiducial inference (see, e.g., Zabell, 1992; Liu and Martin, 2015, and references therein for more discussions on fiducial) is precisely the confidence distribution.*

EXAMPLE 1 (INFERENCE ON THE UNKNOWN MEAN OF A NORMAL DISTRIBUTION). Suppose that we have observed a sample of size n, $y := (y_1, ..., y_n)$, from the normal distribution $N(\theta, 1)$ with unknown $\theta \in \mathbb{R}$. We are interested in making valid inference on the mean parameter θ. Utilizing $T_{y,\theta}$ and its distribution, we have

$$T_{y,\theta} = \ell(y, \hat{\theta}) - \ell(y, \theta) = \frac{1}{2}\left[\sum_{i=1}^{n}(y_i - \hat{\theta})^2 - \sum_{i=1}^{n}(y_i - \theta)^2\right] = -\frac{n}{2}(\bar{y} - \theta)^2, \tag{5}$$

where $\hat{\theta} = \bar{y}$, the sample mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. This implies that $T_{Y,\theta} \sim -\frac{1}{2}\chi_1^2$, as the sampling distribution of $\bar{y}$ is $N(\theta, 1/n)$. Represent the standard normal distribution as $Z \sim N(0, 1)$ and symbolize its cumulative distribution function as $\Phi(\cdot)$. For $Y \sim \mathbf{P}_\theta$,

$$\begin{aligned} F_\theta(t) = \mathrm{P}_{Y|\theta}\left(T_{Y,\theta} \le t\right) &= \mathrm{P}\left(\chi_1^2 \ge -2t\right) \\ &= 2\Phi(-\sqrt{-2t}), \qquad t \le 0. \end{aligned}$$

Note that $\{\theta : \ F_\theta(T_{y,\theta}) \ge \alpha\}$ gives exactly the same classical $(1-\alpha)\%$ z-interval for θ.

By applying the reverse operation, we obtain the distribution $\theta_* \sim N(\bar{y}, 1/n)$. We have $F_{\theta_*}(T_{y,\theta_*}) = 2\Phi(-\sqrt{n}|\bar{y} - \theta_*|) \sim 2\Phi(-|Z|)$ where $Z \sim N(0, 1)$. Since $2\Phi(-|Z|) \sim \mathrm{Uniform}(0, 1)$, the distribution of θ_* satisfies (4), and thus is the confidence distribution of θ. In other words, any $100(1-\alpha)\%$ confidence set from $N(\bar{y}, 1/n)$ has at least $1-\alpha$ chance to cover the true parameter θ.

3. The Calibrated Bootstrap Method

In this section, we propose a computationally feasible method to perform valid parametric inference via an adaptive resampling procedure. This method, termed *calibrated bootstrap* (CB), consists of two steps: *resampling approximation* (RA) and *distributional resampling* (DR). The RA step searches for a resampling scheme that can best approximate the exact confidence region of $\theta \in \Theta$ for each of a set of prespecified confidence coefficients. The DR step selects samples from the bootstrapped estimates obtained in the RA step to construct the confidence distribution of the unknown parameter of interests. These two steps are summarized as two algorithms and are discussed below in Sections 3.1 and 3.4.

3.1. Resampling Approximation

It is seen in Section 2.2 that the key to producing valid inference based on $T_{y,\theta}$ is to obtain the confidence distribution $\theta_* \sim G_y$ that satifies (4). For the general case, in the CB context, we are interested in a bootstrapped sample (or, in theory, population) of estimates $\tilde{\Theta}$, represented in terms of the empirical distribution $\tilde{G}_y$ based on $\tilde{\Theta}$ and

referred to as *bootstrapped confidence distribution* in the sequel, as long as it provides valid inference in the sense:

$$\hat{\theta}_* \sim \tilde{G}_y \quad s.t. \quad F_{\hat{\theta}_*}(T_{y,\hat{\theta}_*}) \sim \text{Uniform}(0,1). \tag{6}$$

It appears difficult to design a resampling scheme to construct a $\tilde{\Theta}$ satisfying (6). Here, we consider to find $\tilde{\Theta}_\alpha$ with the corresponding empirical distribution $\tilde{G}_y^\alpha$ such that for random variable $\hat{\theta}_*$ and distribution function $\tilde{G}_y^\alpha$ that depends on given α and y, when $\hat{\theta}_* \sim \tilde{G}_y^\alpha$,

$$\mathrm{P}_{\hat{\theta}_*}\left(F_{\hat{\theta}_*}(T_{y,\hat{\theta}_*}) \leq \alpha\right) = \alpha, \quad \text{for some pre-defined } \alpha \in (0,1). \tag{7}$$

This is an easier alternative to finding $\tilde{\Theta}$ as (6) entails (7). The following Lemma suggests that when $\hat{\theta}_* \sim \tilde{G}_y^\alpha$, the obtained $1-\alpha$ confidence region using the α quantile of $F_{\hat{\theta}_*}(T_{y,\hat{\theta}_*})$ is exact.

LEMMA 1. *Suppose $T_{Y,\theta}$ is a continuous random variable as a function of $Y \sim \mathbf{P}_\theta$ for $\theta \in \Theta$. For a distribution $\tilde{G}_y^\alpha$ such that when $\hat{\theta}_* \sim \tilde{G}_y^\alpha$, (7) holds for some pre-defined $\alpha \in (0,1)$. Denote the distribution of $F_{\hat{\theta}_*}(T_{y,\hat{\theta}_*})$ as Q_y. Suppose that $\tilde{G}_y^\alpha$ has non-zero density at any $\theta_0 \in \{\theta : F_\theta(T_{y,\theta}) \geq \alpha\}$ and zero density at any $\theta_0 \notin \Theta$. Then, the set*

$$\{\theta' : F_{\theta'}(T_{y,\theta'}) \geq Q_y^{-1}(\alpha)\}, \tag{8}$$

where each θ' represents a realization of the random variable $\hat{\theta}_*$ simulated from $\tilde{G}_y^\alpha$, provides an exact $1-\alpha$ confidence region for θ as defined in Theorem 1, where $Q_y^{-1}(\cdot)$ denotes the inverse function of Q_y.

PROOF. Condition (7) implies that $Q_y^{-1}(\alpha) = \alpha$. Since $\tilde{G}_y^\alpha$ has non-zero density at any $\theta_0 \in \{\theta : F_\theta(T_{y,\theta}) \geq \alpha\}$, (8) is equivalent to $\{\theta : F_\theta(T_{y,\theta}) \geq \alpha\}$. The exactness follows from Theorem 1.

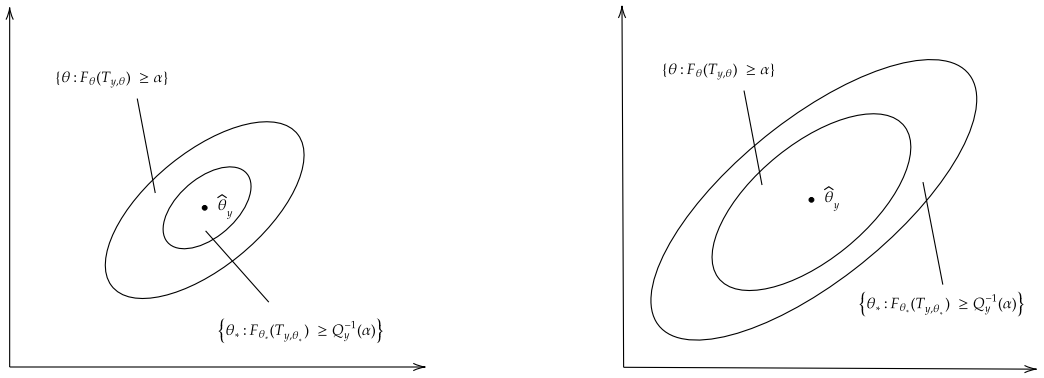
Denote by $r(\cdot)$ a resampling scheme $r(\cdot)$ that is used to generate the resampled data set $\tilde{y} \sim r(y)$. The variability of the estimates introduced by $r(\cdot)$ is captured through the distribution of bootstrap estimates, denoted by $\hat{\theta}_*$. Here, $\hat{\theta}_*$ is defined as the estimated parameter that minimizes the loss function for a given resampled dataset $\tilde{y}$:

$$\hat{\theta}_* = \arg\min_\theta \ell(\tilde{y}, \theta), \quad \tilde{y} \sim r(y) \tag{9}$$

Ideally, for valid inference, the variability of the estimates of θ in (9) should match its theoretically quantified uncertainty in (6) or (7). To search for such a resampling scheme, here we consider a more general method of m-out-of-n bootstrap (Bickel and Sakov, 2008). That is,

$$r(y) = \{\tilde{y}_1, \ldots, \tilde{y}_m\}, \qquad m \geq 1 \tag{10}$$

with $\tilde{y}_i$ $(i = 1, \ldots, m)$ being a random draw with replacement from $\{y_1, ..., y_n\}$. Note that (10) reduces to the standard bootstrap when $m = n$. We take m here to be adaptive to specific models and observations in order to control the variability of the estimates of θ



(a) $m > m_\alpha$ **(b)** $m < m_\alpha$

Fig. 1. Illustration of the over-disperseness of $\hat{\theta}_*$ with different choice of m. The distribution of $F_{\hat{\theta}_*}(T_{y,\hat{\theta}_*})$ for $\hat{\theta}_*$ obtained by the m-out-of-n bootstrap is denoted by Q_y, and its inverse function is denoted by $Q_y^{-1}(\cdot)$. The optimal m denoted by m_α is the m such that the confidence region $\{\hat{\theta}_* : F_{\hat{\theta}_*}(T_{y,\hat{\theta}_*}) \geq Q_y^{-1}(\alpha)\}$ closely matches the desired confidence region $\{\theta : F_\theta(T_{y,\theta}) \geq \alpha\}$.

in (9). Numerically, as $m \to \infty$, we have $\hat{\theta}_* \xrightarrow{p} \hat{\theta}_y$, and $F_{\hat{\theta}_*}(T_{y,\hat{\theta}_*}) \xrightarrow{p} 1$, where $\xrightarrow{p}$ denotes convergence in probability. Theoretically, as a matter of fact, an asymptotic normality result similar to that for the maximum likelihood estimator (see, *e.g.*, Lehmann, 1991, p. 415) can be established by treating the $\hat{\theta} = \arg\min_{\theta\in\Theta} \ell(y, \theta)$ as the true parameter with respect to resampling, because the same proof for the maximum likelihood estimator can go through here due to the fact that the expectation of the score function at $\theta = \hat{\theta}$ is zero. Conversely, as m decreases, $\hat{\theta}_*$ diverges from $\hat{\theta}_y$, causing the distribution of $F_{\hat{\theta}_*}(T_{y,\hat{\theta}_*})$ to gradually shift toward 0. The effect of m on the over-disperseness of the estimates $\hat{\theta}_*$ is illustrated in Figure 1.

For theoretical justifications, here we make a mild assumption that the inference problem is *bootstrap ϵ-calibratable* at level α. Specifically, for fixed y, there exists a positive integer m such that the m-out-of-n bootstrapped estimates $\hat{\theta}_*$ satisfy

$$\alpha \leq \mathrm{P}_{\hat{\theta}_*}\left(F_{\hat{\theta}_*}(T_{y,\hat{\theta}_*}) \leq \alpha\right) \leq \alpha + \epsilon, \tag{11}$$

for some $\epsilon \in [0, \alpha]$ is the tolerance. This assumption ensures the existence of an approximate solution to (7). Additional insights of such a property are given in Supplementary S.2.3. The assumption is considered mild for some common choices of α such as $\{0.05, 0.10, \ldots, 0.95\}$ due to the well-documented flexibility and efficiency of the adaptive m-out-of-n bootstrap in the literature (Bickel and Sakov, 2008; Chakraborty et al., 2013; Jiang and Liu, 2024). Similarly, we also make another mild assumption that the set of the finite number of possible values provided by the m-out-of-n bootstrap is dense enough for practically accurate inference on θ given y; see the discussion in Section 5 on the use of weighted likelihood estimation as alternatives to or a generalization of resampling in the context of bootstrapping.

In the first step of the proposed CB method, our primary objective is to determine

an optimal value of m, such that the empirical distribution of the obtained bootstrap estimates of θ in (9) from the m-out-of-n bootstrap best approximates the distribution of $\tilde{G}_y^{\alpha}$ in (7). This can be done using the *stochastic approximation* (SA) algorithm (Robbins and Monro, 1951). See Syring and Martin (2019) for the application of SA for obtaining a specific target coverage level in a different context. For an unbiased estimator of the distribution of $F_{\hat{\theta}_*}(T_{y,\hat{\theta}_*})$ that is needed for a SA implementation, we consider estimating the distribution of $F_{\hat{\theta}_*}(T_{y,\hat{\theta}_*})$ with the empirical distribution function obtained by B resampling repetitions. Suppose that after B resampling repetitions, we have B simulated values

$$\begin{aligned} U_1 &= F_{\hat{\theta}_*^{(1)}}(T_{y,\hat{\theta}_*^{(1)}}) \\ &\vdots \\ U_B &= F_{\hat{\theta}_*^{(B)}}(T_{y,\hat{\theta}_*^{(B)}}), \end{aligned}$$

where $\hat{\theta}_*^{(b)}$ $(b = 1, \ldots, B)$ denotes the obtained estimation of θ by minimizing the loss functions with regard to the b-th bootstrap samples. For sufficiently large B, objective (7) is equivalent to

$$\lim_{B\to\infty} \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}(U_b \leq \alpha) = \alpha \tag{12}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. The approximation of $F_{\hat{\theta}_*}(T_{y,\hat{\theta}_*})$ for each $\hat{\theta}_*$ can be done by sampling N new observation vectors $y^{(1)}, \ldots, y^{(N)}$ from the population $\mathbf{P}_{\hat{\theta}_*}$

$$F_{\hat{\theta}_*}(T_{y,\hat{\theta}_*}) \approx \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(T_{y^{(i)},\hat{\theta}_*} \leq T_{y,\hat{\theta}_*}\right). \tag{13}$$

Alternative approaches, such as by reweighting the observed data using the SIR approach of Rubin (1987), can also be considered.

The right-hand side of (13) is seen as an unbiased estimator of $F_{\hat{\theta}_*}(T_{y,\hat{\theta}_*})$. Making use of this fact, we propose a computationally efficient SA method of finding the optimal m for a pre-specified value of $\alpha \in (0, 1)$ subject to constraints (12). The method is summarized as Algorithm 1. The step size c used in the algorithm controls the convergence speed, and one practical choice is $c = d \cdot n$ for some $d > 0$. In the simulation study in Section 4, we use $d = 10$. Additionally, in the illustrative examples presented in this paper, we use a small number of resampling repetitions, specifically $B = 10$. This approach has yielded satisfactory convergence results. The following theorem provides necessary theoretical results on the required stochastic update in Algorithm 1. The proof is given in the Supplementary S.2.

THEOREM 3. *For a given targeted coverage probability* $1 - \alpha$ $(0 < \alpha < 1)$*, let the observed coverage probability be defined as a function of* m*:*

$$f^{\alpha}(m) \coloneqq \mathrm{P}_{\hat{\theta}_*(m)}\left(F_{\hat{\theta}_*(m)}(T_{y,\hat{\theta}_*(m)}) \leq \alpha\right)$$

Algorithm 1: The Resampling Approximation algorithm

Specify the target coverage level α and choose the number of bootstrap replications B;
Initialize $m^{(0)}$ with $m^{(0)} = n$ and the iteration number t with $t = 0$;
while $m^{(t)}$ not converged **do**
    Let $m = \lfloor m^{(t)} \rfloor + \text{Bernoulli}(m^{(t)} - \lfloor m^{(t)} \rfloor)$, where $\lfloor m^{(t)} \rfloor$ denotes the largest integer not greater than $m^{(t)}$ and Bernoulli(p) a Bernoulli random variable with parameter p;
    Sample $\{(x_i^*, y_i^*) : i = 1, ..., m^{(t)}\}$ or, in the matrix notation, (X^*, Y^*) from $\{(x_i, y_i) : i = 1, ..., n\}$ with replacement;
    Compute $\hat{\theta}_* = \arg\min_\theta \ell(\theta, X^*, Y^*)$;
    Evaluate $\ell^{(t)} = -[\ell(\hat{\theta}_*, X, Y) - \ell(\hat{\theta}, X, Y)]$;
    Initialize $P = 0$;
    for $b \leftarrow 1$ **to** B **do**
        Sample $\{(x_i^{**}, y_i^{**}) : i = 1, ..., n\}$ or, in the matrix notation, (X^{**}, Y^{**}) from model $\mathbf{P}_{\hat{\theta}_*}$;
        Compute $\hat{\theta}_{**} = \arg\min_\theta \ell(\theta, X^{**}, Y^{**})$;
        Evaluate $S^{(b)} = -[\ell(\hat{\theta}_*, X^{**}, Y^{**}) - \ell(\hat{\theta}_{**}, X^{**}, Y^{**})]$;
        if $S^{(b)} \leq \ell^{(t)}$ **then**
            $P \leftarrow P + 1$;
    Set $Z_t \leftarrow \frac{\mathbb{I}\{P/B \leq \alpha\} - \alpha}{\sqrt{B\alpha(1-\alpha)}}$;
    Update $m^{(t+1)} = m^{(t)} + \frac{c}{t+1} Z_t$, where c is a predefined constant;
    Set $t \leftarrow t + 1$;
Return $\lfloor m^{(t)} \rfloor - 1$.

where $\hat{\theta}_*(m)$, with the resulting distribution $\mathrm{G}_{y,m}$, is the bootstrap estimate of θ obtained with the m-out-of-n bootstrap. The convergence of Algorithm 1 to the desired solution

$$m^*_\alpha = \min_{m\geq 1}\{f^\alpha(m) - \alpha : f^\alpha(m) \geq \alpha\}$$

is guaranteed with probability one if $f^\alpha(m)$ is non-increasing with respect to m and

$$|f^\alpha(m_0) - \alpha| > |f^\alpha(m^*_\alpha) - \alpha| \tag{14}$$

for all $m_0 \neq m^*_\alpha$, under the regularity conditions of likelihood function given in Supplementary S.2. Moreover, if the problem is bootstrap ϵ-calibratable by (11), the constructed confidence interval by (8) is at least at the $1-\alpha$ level.

COROLLARY 1. *The sufficient conditions of non-increasing $f^\alpha(m)$ in Theorem 3 are:*

C1. For θ_1, θ_2 such that $\ell(y,\theta_1) \geq \ell(y,\theta_2)$, we have $F_{\theta_1}(T_{y,\theta_1}) \leq F_{\theta_2}(T_{y,\theta_2})$, and

C2. For any $m_1 > m_2 > 0$, $\ell(y,\hat{\theta}_*(m_2))$ first-order stochastically dominates $\ell(y,\hat{\theta}_*(m_1))$.

Moreover, Theorem 3 holds asymptotically as $m, n \to \infty$, given that the asymptotic normality of maximum likelihood estimator (MLE) holds.

REMARK 4. *One sufficient condition for C1 in Corollary 1 to hold true is the distribution of $T_{Y,\theta}$ being independent of θ for $Y \sim \mathrm{P}_\theta$. Further insights about this sufficient condition are elaborated in Martin (2015) as well as Supplementary S.3. It can also be verified that condition C1 is satisfied in commonly used models, such as the linear regression model. Condition C2, on the other hand, is grounded in the concept that the model loss decreases as the sample size increases. The demonstration of the example of linear regression model satisfying these conditions are given in Supplementary S.3.2.*

Now, we attempt to approximate the confidence distribution $\hat{\theta}_* \sim \tilde{G}$ in (6) with some resampling scheme. Note that the sufficient condition for (6) is that for all $\alpha \in (0,1)$, (7) holds true. This motivates us to use the SA method to find the optimal m for each in a range of target coverage probabilities, such as $0.05, 0.15, ..., 0.95$. The estimated resampling scheme is the m-out-of-n bootstrap with a mixture of these m values. Note that due to the limitation in the versatility of m-out-of-n bootstrap, the confidence distribution may not be perfectly recovered. In this case, further refinement methods, as will be dicussed in Section 3.4, are necessary.

3.2. RA with a Simple Example

Consider the example in Section 2.2 on the inference of the mean parameter θ with y, a sample containing n observations from the model $Y \sim N(0,1)$. Here we illustrate the application of RA to numerically approximate the $(1-\alpha)\%$ z-interval for θ with some pre-specified α.

By (5), for each bootstrap estimate $\hat{\theta}^* \in \Theta$, the Monte-Carlo estimated function value $F_{\hat{\theta}_*}(T_{y,\hat{\theta}_*})$ with B repetitions takes the form

$$F_{\hat{\theta}_*}(T_{y,\hat{\theta}_*}) \approx \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\left(T_{y^{(b)},\hat{\theta}_*} \le T_{y,\hat{\theta}_*}\right)$$
$$= \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\left(\frac{1}{2}\sum_{i=1}^{n}(\bar{y}^{(b)} - y_i^{(b)})^2 - \frac{1}{2}\sum_{i=1}^{n}(\hat{\theta}_* - y_i^{(b)})^2 \le \frac{1}{2}\sum_{i=1}^{n}(\bar{y} - y_i)^2 - \frac{1}{2}\sum_{i=1}^{n}(\hat{\theta}_* - y_i)^2\right) \tag{15}$$

where $y^{(b)}$ is a sample of size n from the model $N(\hat{\theta}_*, 1)$.

In the RA process, an initial value $m^{(0)}$ is used to conduct the m-out-of-n bootstrap on the observed data y to obtain the resample data $\tilde{y}$, which yields the parameter value $\hat{\theta}_* = \bar{\tilde{y}}$, the MLE of the resampled data. Next, the function value $F_{\hat{\theta}_*}(T_{y,\hat{\theta}_*})$ is calculated by (15), based on which we update $m^{(0)}$ with

$$m^{(t+1)} = m^{(t)} + \frac{c}{t+1}\frac{\mathbb{I}\{F_{\hat{\theta}_*}(T_{y,\hat{\theta}_*}) \le \alpha\} - \alpha}{\sqrt{B\alpha(1-\alpha)}},$$

taking the iteration number as $t = 0$, where $c = 10 \cdot n$ is a predefined constant. Finally, we increase the iteration number t by 1 and repeat the whole process for the repetitive updates of m until convergence.

With the obtained m from the RA step, we can then conduct the m-out-of-n bootstrap to obtain the resampled $\hat{\theta}_*$ whose distribution satisfies $\mathrm{P}_{\hat{\theta}_*}\left(F_{\hat{\theta}_*}(T_{y,\hat{\theta}_*}) \le \alpha\right) \approx \alpha$. By Theorem 1, we wish to construct the $1-\alpha$ confidence interval $\{\hat{\theta}_* : F_{\hat{\theta}_*}(T_{y,\hat{\theta}_*}) \ge \alpha\}$ with the set of $\hat{\theta}_*$ values obtained from the m-out-of-n bootstrap. It can be seen from (15) that the distribution of the left hand side of the inequality does not depend on $\hat{\theta}_*$. As a result, for θ_1, θ_2 such that $\ell(y, \theta_1) \ge \ell(y, \theta_2)$, we have $F_{\theta_1}(T_{y,\theta_1}) \le F_{\theta_2}(T_{y,\theta_2})$. Thus, the $1-\alpha$ confidence interval can be obtained by the range of the $\hat{\theta}_*$ values for which the corresponding loss function values fall within the lowest $1-\alpha$ quantile of all obtained loss function values.

For numerical evaluations, we experimented with the case $n = 50, \theta = 1, \alpha = 0.05$. The theoretical interval is $(0.757, 1.312)$ and the proposed RA method gives $(0.756, 1.313)$.

3.3. RA with a Linear Regression Example

To illustrate the proposed method, here we present a study on linear regression under a high-dimensional setting. Further applications of Algorithm 1 are illustrated in Section 4.

Example 2 (High-Dimensional Linear Regression). Consider the linear regression model

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim N_n(0, \sigma^2 I), \beta \in \mathbb{R}^p, \tag{16}$$

where X is the $(n \times p)$ full-rank design matrix, Y is the vector of n observed responses, $1 \le p < n$, and $\sigma^2 > 0$ denotes the variance of the noise term. In the context of high-dimensional linear regression, one may imagine the scenario where $p/n \to c$ for some $0 < c < 1$ as $n \to \infty$.

Table 1. Upper bounds of the constructed confidence Interval for $(\beta - \hat{\beta})'(X'X)(\beta - \hat{\beta})/p$ at various significance levels α, comparing the standard bootstrap and CB. Each method uses 1,000 bootstrap samples, with the number of observations m resampled from the original dataset (size $n = 500$), as determined in the RA step, specified in brackets. All intervals have lower bounds fixed at 0.

	0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
Truth	1.197	1.120	1.075	1.041	1.010	0.981	0.952	0.920	0.881	0.818
Bootstrap	1.820	1.650	1.547	1.454	1.385	1.324	1.262	1.190	1.107	0.992
	[500]	[500]	[500]	[500]	[500]	[500]	[500]	[500]	[500]	[500]
CB	1.176	1.111	1.074	1.032	1.006	0.984	0.938	0.919	0.896	0.801
	[651]	[630]	[617]	[611]	[603]	[597]	[598]	[584]	[569]	[557]

Suppose that the estimand of interest is the unknown vector β of regression coefficients. Theoretical results from the fiducial inference perspective or, more precisely, the CIM (Conditional Inferential Models) perspective (see, *e.g.*, Martin and Liu, 2015) shows that for the estimates

$$\hat{\beta} = (X'X)^{-1}X'Y, \quad \hat{\sigma} = \frac{1}{n-p}Y'\left(I - X(X'X)^{-1}X'\right)Y,$$

the fiducial or confidence distribution of β is given by

$$\beta^* \sim t_p\left(\hat{\beta}, \hat{\sigma}^2(X'X)^{-1}, n-p\right), \tag{17}$$

where $t_p(\cdot)$ denotes the p-dimensional multivariate student-t distribution, and if σ is known,

$$\beta^* \sim N_p\left(\hat{\beta}, \sigma(X'X)^{-1}\right) \tag{18}$$

where $N_p(\cdot)$ denotes the p-dimensional multivariate Gaussian distribution. Apparently, (17) and (18) can be used to construct the classical valid frequentist confidence region for β. More relevant details and discussion, focusing on the high-dimensional setting, are provided in Supplementary S.5 and S.6.

For a simulation study, we took $n = 500$ and $\kappa = p/n = 0.3$, a case in the context of bootstrap for high-dimensional problems as discussed in El Karoui and Purdom (2018). We set $\sigma = 1$ to be known. Our goal is to conduct a valid and efficient joint inference on β through resampling. Our numerical experiments indicate that both standard bootstrap and the residual bootstrap (Freedman, 1981) yield unsatisfactory results in cases where $\kappa = p/n = 0.3$, with variations in the choice of n (see Supplementary S.7 for details). This observation aligns with the results reported in Bickel and Freedman (1983) and El Karoui and Purdom (2018). Here, we show that the proposed CB method can give a valid and efficient joint inference on β. The SA algorithm to find the optimal m was applied to each of 10 equally spaced target coverage probabilities $0.05, 0.15, ..., 0.95$. The trajectories of the 10 SA runs and the corresponding 10 estimated m-values are shown in Figure 2 (a). We see that all estimated m values are larger than n in this case. We compare the constructed frequentist $1-\alpha$ confidence intervals for $(\beta-\hat{\beta})'(X'X)(\beta-\hat{\beta})/p$ with standard bootstrap and the estimated m-out-of-n bootstrap by RA with the 10 confidence coefficients α. Both bootstrap confidence intervals are constructed by taking

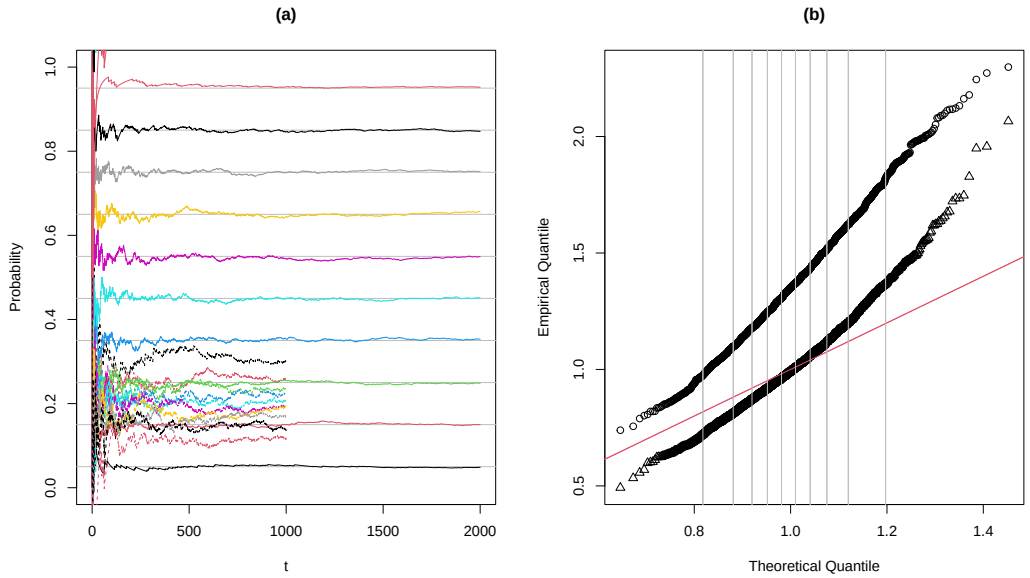


Fig. 2. Illustration of SA on the Gaussian linear regression with unit error variance, $n = 500$, $\kappa = p/n = 0.3$, a case in the context of bootstrap for high-dimensional problems in El Karoui and Purdom (2018). Plot (a) shows the trajectories of 10 SA runs for each of 10 equally-spaced targeted coverage probabilities $0.05, 0.15, ..., 0.95$. The trajectories of 10 m-values in terms of m/n and $m/n - 1$ are displayed in dotted lines (compressed to half its original length by displaying the values every two iterations for clarity). The solid lines denote the cumulative mean values of $F_\theta(T_{y,\theta})$. The circle "$\circ$" points in Plot (b) is the Q-Q plot of 1,000 bootstrap values of $(\beta_* - \hat{\beta})'(X'X)(\beta_* - \hat{\beta})/p$ against the theoretical quantiles. The triangle "$\triangle$" points show the corresponding quantiles for those obtained by bootstrap with the 10 estimated m values. The gray vertical lines denote the 10 quantiles correspond to the 10 targeted coverage probabilities.

0 as the lower bound and $1-\alpha$ quantile of the resampled values $(\tilde{\hat{\beta}}-\hat{\beta})'(X'X)(\tilde{\hat{\beta}}-\hat{\beta})/p$ as the upper bound. As a matter of fact, similar to the reasoning in Section 3.2, it can be shown that in this example, this interval construction method for the proposed RA is equivalent to (8). The results are summarized in Table 1. It can be seen that our proposed method gives an estimate numerically very close to the true confidence region derived theoretically.

Figure 2 (b) compares the distribution of the standard bootstrap estimates of the confidence distribution $(\beta_*-\hat{\beta})'(X'X)(\beta_*-\hat{\beta})/p$ against the theoretical distribution obtained from (18). The corresponding distribution obtained by the m-out-of-n bootstrap with the mixture of 10 estimated values of m (each with equal probabilities) is also shown. It can be seen that the resampling scheme found by RA dramatically outperforms the standard bootstrap in recovering the theoretical distribution of $(\beta_*-\hat{\beta})'(X'X)(\beta_*-\hat{\beta})/p$. However, it appears challenging to perfectly recover the desired theoretical distribution via a mixture of the m-out-of-n bootstrap, regardless of the number of mixture components. This suggests the potential for exploring alternative resampling schemes that can provide greater flexibility than the m-out-of-n bootstrap. Furthermore, this leads to our proposed refinement method that is discussed next in Section 3.4.

3.4. Distributional Resampling of $T_{y,\theta}$

As elaborated in Section 3.1, to approximate the confidence distribution of θ with bootstrapped $\hat{\theta}_*$'s, it requires that (6) holds. This motivates a simple refinement method to resample the bootstrapped estimates $\hat{\theta}_*$'s in (9) obtained in the RA step to create a new distribution such that the values of $F_{\hat{\theta}_*}(T_{y,\hat{\theta}_*})$ with the resampled $\hat{\theta}_*$'s closely approximate a sample from the standard uniform distribution. Such resampled distribution can then be seen as an approximation to the targeted confidence distribution and can be used for constructing confidence sets at all $\alpha \in (0,1)$ levels by Theorem 2. The proposed method is summarized into the following three-step algorithm, which is referred to as *distributional resampling* (DR).

Algorithm 2 (Distributional Resampling). *Suppose that a set of m values for the m-out-of-n bootstrap method is obtained by applying the SA algorithm with varying target coverage level α (e.g. $\alpha = \{0.05, 0.50, 0.95\}$). The DR algorithm creates a sample of selected bootstrap estimates in three steps:*

(a) Create B bootstrapped estimates $\hat{\theta}_*^{(1)}, \ldots, \hat{\theta}_*^{(B)}$ with sufficiently large B using the m values obtained by the SA algorithm.

(b) Compute $U_b = F_{\hat{\theta}_*^{(b)}}(T_{y,\hat{\theta}_*^{(b)}})$ with Monte Carlo approximation in (13) for $b = 1, \ldots, B$.

(c) Set $\tilde{\Theta} = \emptyset$ as the set of resampled estimates and repeat the following steps for B times:

(i) Draw $u \sim$ Uniform$(0,1)$;

(ii) Find $b_* = \arg\min_b |U_b - u|$;

(iii) Add $\hat{\theta}_*^{(b_*)}$ to set $\tilde{\Theta}$.

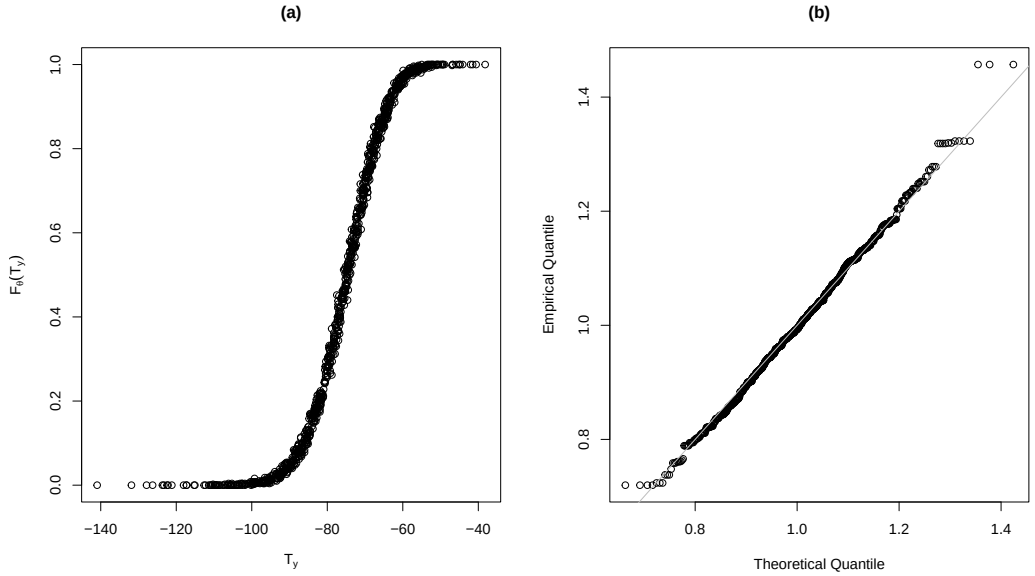


Fig. 3. Illustration of the DR step after RA. Plot (a) shows the estimate $\hat{F}_{\hat{\theta}_*}(.)$ via Monte Carlo with 1,000 $\hat{\theta}_*$ values obtained by bootstrapping with the 10 estimated m values explained in Figure 2. Plot (b) is the Q-Q plot of the bootstrap values of $(\beta_* - \hat{\beta})'(X'X)(\beta_* - \hat{\beta})/p$ selected by DR against the theoretical quantiles.

REMARK 5. *The DR algorithm can be employed with any sets of* $\hat{\theta}_*^{(1)}, \ldots, \hat{\theta}_*^{(B)}$. *Nonetheless, the RA step, which offers a close, albeit not flawless, approximation to the true confidence distribution of* θ, *can produce a more optimal effective sample size, making it a more effective choice. The example in Supplementary S.4 shows that the construction of the confidence distribution via DR can be practically impossible without the RA step. The effect of the varying* α *levels used in the RA step on the final result is also empirically studied in Supplementary S.4, showing that the result is robust to the grid density of* α. *Another example in Martin (2023a) where the true confidence distribution is known is used to demonstrate the efficiency of the approximation in step (c) of the DR algorithm (see Supplementary S.8).*

REMARK 6. *For computational efficiency, the (a) and (b) steps of the DR algorithm can be combined into the SA algorithm in the RA step. Specifically, the bootstrapped estimates* $\hat{\theta}_*^{(b)}$ *as well as the value* $F_{\hat{\theta}_*^{(b)}}(T_{y,\hat{\theta}_*^{(b)}})$ *are collected along the SA procedure. The resulting algorithm is given in Supplementary S.9.*

EXAMPLE 2 (CONT'D). We applied the DR method to the bootstrapped estimates of θ obtained with the m values from the RA process. Figure 3 (a) shows the scatter plot of the initial m-out-of-n bootstrapped values of $F_{\hat{\theta}_*}(T_{y,\hat{\theta}_*})$ versus the corresponding values of $T_{y,\hat{\theta}_*}$. Similar to Figure 2 (b), Figure 3 (b) shows the empirical quantiles of the

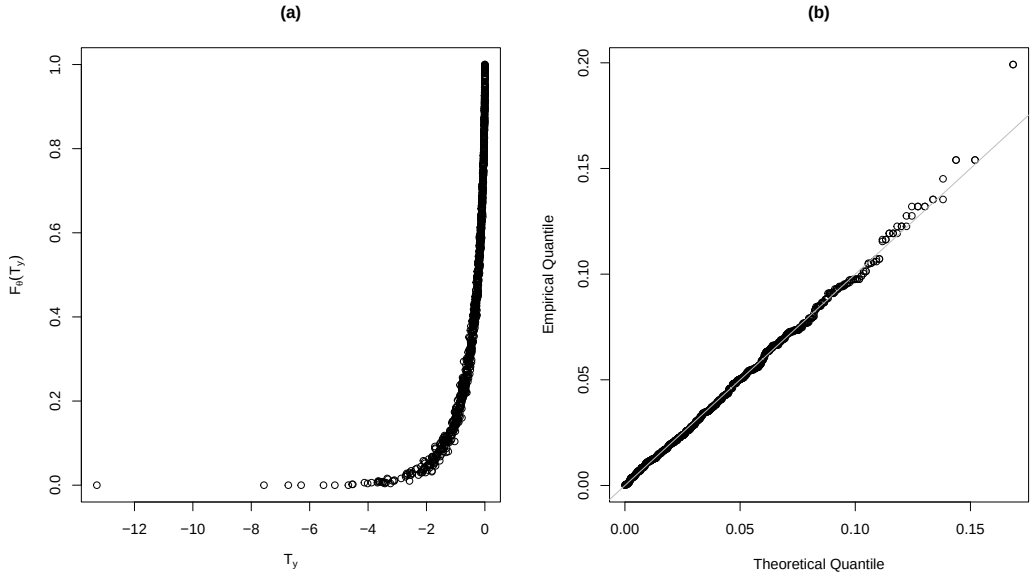


Fig. 4. Illustration of the results after RA and DR for marginal inference of β_1. Plot (a) shows the estimate $\hat{F}_{\hat{\theta}_*}(.)$ via Monte Carlo with 1,000 $\hat{\theta}_*$ values obtained by bootstrapping with the 10 estimated m values through the RA step. Plot (b) is the Q-Q plot of the bootstrap values of $|\beta_1^* - \hat{\beta}_1|$ selected by DR against the theoretical quantiles.

resampled estimates $(\beta_* - \hat{\beta})'(X'X)(\beta_* - \hat{\beta})/p$ obtained with DR versus the theoretical quantiles. Comparing Figure 3 (b) to Figure 2 (b), we see that the application of RA followed by DR creates a nearly perfect approximation to the true confidence distribution.

3.5. Marginal Parametric Inference

In the previous section, it was shown that the proposed RA method followed by the DR refinement process can provide a valid inference on the model parameters $\theta \in \Theta$. In practice, we might only be interested in the marginal inference of some functions of θ. For example, when partitioning the model parameter as $\theta = (\theta_1, \theta_2) \in \Theta = \Theta_1 \times \Theta_2$, we are interested in constructing a confidence region for θ_1, while treating θ_2 as the nuisance parameter. Here, we propose a solution inspired by Martin (2015, 2023b), which replaces the generalized association function (2) with a form of the relative profile likelihood

$$T_{y,\theta_1} = \ell(y, \hat{\theta}_y) - \max_{\theta_2} \ell(y, \theta_1, \theta_2), \qquad \theta \in \Theta, y \in \mathbb{Y}. \tag{19}$$

The idea is to create an association function that depends on θ_1 alone, by marginalizing out θ_2 via the profile likelihood.

In our proposed CB method, we conduct the RA process using the generalized association function defined in (19). If the distribution of T_{Y,θ_1} as a function of $Y \sim \mathbb{P}_\theta$ is independent of θ_2, the exactness of the constructed confidence region for θ_1 via

$\{\theta_1 : F_{\theta_1}(T_{y,\theta_1}) \geq \alpha\}$ for $\alpha \in (0,1)$, as previously elaborated in Theorem 1 is straightforward to establish (see, *e.g.*, Martin, 2015). Even if when T_{y,θ_1} being independent of θ_2 is hard to verify, Martin (2023b) demonstrates that inference conducted using the profile likelihood-based association function (19) remains both theoretically valid and empirically efficient. This robust theoretical foundation further supports the validity of the proposed CB approach.

Example 2 (Cont'd). We considered the marginal inference of the individual coefficient β_1. The RA and DR steps are carried out with the relative profile likelihood (19), where $\theta_1 := \beta_1$ and $\theta_2 := \{\beta_j \,|\, j \neq 1\}$. The results are shown in Figure 4. It can be seen that the CB method provides a very good approximation to the true fiducial distribution of $|\beta_1^* - \hat{\beta}_1|$.

4. Applications in L_1-penalized Linear Regression

Revisit the linear regression model (16) articulated in Example 2. The L_1-penalized estimator of β, also known and popularized as the Lasso estimator (Tibshirani, 1996), is considered here and can be written as

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2}(Y - X\beta)^T(Y - X\beta) + \lambda \sum_{i=1}^{p} |\beta|_i,$$

where $\lambda > 0$ is the tuning parameter. The Lasso estimator is useful especially when $p > n$, as it can provide variable selection by shrinking some coefficients toward zero. The involvement of the regularization term, allowing for a more robust estimator compared to the ordinary least squares (OLS) estimator, can result in better prediction power.

Although Lasso is widely used in real data analysis scenarios such as genome-wide association studies (GWAS, Uffelmann et al., 2021; Zeng et al., 2015), designing a valid and efficient inference procedure is generally considered difficult, due to the added penalty term. Resampling-based approaches, such as bootstrap and permutation tests (Arbet et al., 2017), are the most commonly used methods for model inference. However, previous theoretical studies have shown that even in the asymptotic case, the validity of the bootstrap-based inference procedure for Lasso is difficult to establish (Fu and Knight, 2000; Chatterjee and Lahiri, 2010, 2011).

A remarkable benefit of the proposed CB method lies in its ability to offer a valid frequentist inference procedure for methods like Lasso, where valid inference is challenging to derive mathematically. We will first use a simulation study to demonstrate the efficiency of the proposed method in Section 4.1 and then compare our proposed method with other methods in a real data example in Section 4.2.

4.1. Simulation Study

Similar to the high-dimensional setting in Example 2, our simulation studies consider the scenario where the ratio $\kappa = p/n \to 1$. Specifically, we increase κ by varying its value with $\{0.3, 0.5, 0.9\}$, alongside an increase in sample size n with $\{100, 200, 500\}$. We let $X_{ij} \sim N(0,1)$ and standardize X to have mean 0 and standard deviation 1 in each column. For the vector of true parameters β, we assume the signal to be sparse by setting

Table 2. Estimated coverage probabilities of the true β and expected magnitude for the constructed $100(1-\alpha)\%$ joint confidence region with different values of α using the model settings ($p/n \to 1$). The standard deviation of each value (estimated with bootstrap) is given in parentheses.

		CB		Standard Bootstrap		Residual Bootstrap	
Model	α	Coverage	Magnitude	Coverage	Magnitude	Coverage	Magnitude
$n = 100$, $p = 30$	0.05	0.952 (0.010)	0.463 (0.002)	0.780 (0.019)	0.270 (0.003)	0.958 (0.009)	0.471 (0.002)
	0.15	0.862 (0.015)	0.328 (0.001)	0.604 (0.022)	0.183 (0.002)	0.872 (0.015)	0.334 (0.001)
	0.25	0.780 (0.019)	0.259 (0.001)	0.486 (0.022)	0.142 (0.002)	0.786 (0.018)	0.265 (0.001)
$n = 200$, $p = 100$	0.05	0.946 (0.010)	0.201 (0.001)	0.446 (0.022)	0.077 (0.001)	0.962 (0.009)	0.205 (0.000)
	0.15	0.864 (0.015)	0.152 (0.000)	0.298 (0.020)	0.052 (0.001)	0.864 (0.015)	0.154 (0.000)
	0.25	0.758 (0.019)	0.125 (0.000)	0.192 (0.018)	0.040 (0.001)	0.762 (0.019)	0.127 (0.000)
$n = 500$, $p = 450$	0.05	0.950 (0.010)	0.045 (0.000)	0.844 (0.016)	0.034 (0.000)	0.946 (0.010)	0.046 (0.000)
	0.15	0.866 (0.015)	0.034 (0.000)	0.726 (0.020)	0.026 (0.000)	0.876 (0.015)	0.035 (0.000)
	0.25	0.762 (0.019)	0.028 (0.000)	0.646 (0.021)	0.023 (0.000)	0.778 (0.019)	0.029 (0.000)

$\beta = (3, 0, \ldots, 0)$. In our data simulation, we set $\sigma = 1$. When performing inference with CB, both β and σ are treated as unknown. In the case of Lasso, a key concern is the valid inference of β. We consider both the joint inference on β and the marginal inference on β_1 (the true signal). To compare with other bootstrap-based methods, we include the standard bootstrap (Efron, 1979) as well as the (debiased) residual bootstrap (Chatterjee and Lahiri, 2010). Specifically, carefully designed residual bootstrap has been shown to be particularly effective in scenarios where $p/n \to 1$ (Lopes, 2014), offering a powerful alternative to the standard bootstrap.

To perform the proposed CB, we take $\pi(\theta) = \lambda \sum_{i=1}^{p} |\beta|_i$ in (1) to formulate the association function $T_{y,\theta}$ with penalized likelihood. The penalty term λ used for regularization takes values of $\lambda \in \{20.1, 40.2, 63.1\}$ correspondingly for the three cases, which is initially determined by 10-fold cross-validation on the entire observed dataset, and remains constant across all 500 repetitions. We fix σ^2 at $\hat{\sigma}^2 = \frac{(\beta-\hat{\beta}_y)'(X'X)(\beta-\hat{\beta}_y)}{n-\sum_{i=1}^{p} \mathbb{I}\{\hat{\beta}_y \neq 0\}}$, a refined estimator given in Reid et al. (2016), where $\hat{\beta}_y$ denotes the Lasso estimates of β on the entire observed dataset. The implementation of CB uses the R package `glmnet` (Friedman et al., 2010) and `natural` (Yu and Bien, 2019).

A valid joint inference aims to provide a close approximation of the true distribution of the quantity $(\beta-\hat{\beta}_y)'(X'X)(\beta-\hat{\beta}_y)/p$. We apply the RA step followed by the DR step the create an approximate confidence distribution for the parameter β, and construct the $1-\alpha$ confidence region $\{\beta : (\beta-\hat{\beta}_y)'(X'X)(\beta-\hat{\beta}_y)/p \leq q_{1-\alpha}\}$, where $q_{1-\alpha}$ is determined in such a way that $100(1-\alpha)\%$ of the samples $\hat{\beta}_*$ obtained from the DR step fall within the region. We refer to $q_{1-\alpha}$ as the magnitude of the confidence region, analogous to the length of the confidence interval in a multidimensional context. The summarized results, presented in Table 2, demonstrate that CB can achieve the desired joint coverage rate while the standard bootstrap exhibits significant undercoverage. The residual bootstrap also performs effectively in such $p/n \to 1$ simulation setting, as validated in Lopes (2014); however, it tends to yield more conservative intervals than CB.

We also performed the marginal inference on β_1. The confidence interval is constructed as $\{\beta_1 : |\beta_1 - \hat{\beta}_{y,1}| \leq q_{1-\alpha}\}$ where $q_{1-\alpha}$ is chosen such that $100(1-\alpha)\%$ of the samples $\hat{\beta}_1^*$ obtained from the DR step are covered by the interval. The summarized results are shown in Table 3, demonstrating that CB can achieve the desired marginal coverage rate. In contrast, the standard bootstrap method continues to exhibit significant undercoverage.

Table 3. Estimated coverage probabilities of the true β_1 and expected length for the constructed $100(1-\alpha)\%$ marginal confidence interval with different values of α using the model settings ($p/n \to 1$). The standard deviation of each value (estimated with bootstrap) is given in parentheses.

		CB		Standard Bootstrap		Residual Bootstrap	
Model	α	Coverage	Length	Coverage	Length	Coverage	Length
$n = 100$, $p = 30$	0.05	0.950 (0.010)	0.735 (0.002)	0.510 (0.022)	0.409 (0.002)	0.950 (0.010)	0.739 (0.001)
	0.15	0.844 (0.016)	0.615 (0.002)	0.320 (0.021)	0.298 (0.001)	0.860 (0.016)	0.616 (0.001)
	0.25	0.764 (0.019)	0.544 (0.001)	0.212 (0.018)	0.238 (0.001)	0.756 (0.019)	0.544 (0.001)
$n = 200$, $p = 100$	0.05	0.940 (0.011)	0.634 (0.001)	0.218 (0.018)	0.287 (0.001)	0.958 (0.009)	0.639 (0.001)
	0.15	0.858 (0.016)	0.549 (0.001)	0.072 (0.012)	0.210 (0.001)	0.862 (0.015)	0.553 (0.000)
	0.25	0.746 (0.019)	0.499 (0.001)	0.038 (0.009)	0.167 (0.001)	0.756 (0.019)	0.501 (0.000)
$n = 500$, $p = 450$	0.05	0.942 (0.010)	0.397 (0.001)	0.188 (0.017)	0.178 (0.000)	0.938 (0.011)	0.400 (0.000)
	0.15	0.852 (0.016)	0.344 (0.000)	0.086 (0.013)	0.131 (0.000)	0.862 (0.015)	0.345 (0.000)
	0.25	0.750 (0.019)	0.313 (0.000)	0.048 (0.010)	0.104 (0.000)	0.760 (0.019)	0.313 (0.000)

4.2. Real Data Example

Consider the diabetes study introduced by Efron et al. (2004). This study encompasses ten baseline variables ($p = 10$), including age, sex, body mass index (BMI), mean arterial pressure (MAP), and six blood serum measurements (S1-S6). The primary response variable of interest corresponds to a quantitative measure of disease progression, recorded one year after baseline assessment, for each of the $n = 442$ diabetes patients in the dataset.

We start by standardizing the variables so that $\sum_{i=1}^{n} x_{ij} = 0, \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1,$ for $j = 1, \ldots, p$ and $\frac{1}{n}\sum_{i=1}^{n} y_i = 0$. We also performed a routine linear regression diagnostic analysis to verify that the basic assumptions of the linear regression model are appropriate. Then, we fit a Lasso version of the model to the dataset for predicting the response variable of interest, perform variable selection, and make inference on regresion coefficients of the predictors. The penalty value $\lambda \approx 520$ is selected by 10-fold cross-validation, and σ^2 is estimated using the residual sum of squares from the standard linear regression model including all predictors.

Lasso yields a model with seven variables: sex, BMI, MAP, S1, S3, S5 and S6. For these selected variables, we compare the confidence intervals obtained by our proposed CB method with those obtained by the standard bootstrap. We also include a recent method of Lee et al. (2016) designed for post-selective inference of Lasso. The method, which we refer to as exact POSI, is capable of constructing a valid confidence interval for the predictors selected by Lasso, while conditioning on the selected model. The constructed 95% intervals with the three methods are shown in Figure 5. For the three methods, our method ensures a minimum of 95% coverage under the full model with 10 predictors, in contrast to exact POSI, which guarantees this level of coverage conditional on the model with seven selected predictors. The standard bootstrap does not warrant 95% coverage. The intervals from our CB method are shown to be comparable in length to those of other methods but are notably shorter than the exact POSI for S6. The results also indicate a consensus among all methods regarding the non-significance of S6.

5. Concluding Remarks

In this paper, we proposed a resampling approximation approach that enables valid finite sample joint inference based on likelihood functions and marginal inference based on profile likelihood. It can be easily extended to cases where general loss functions

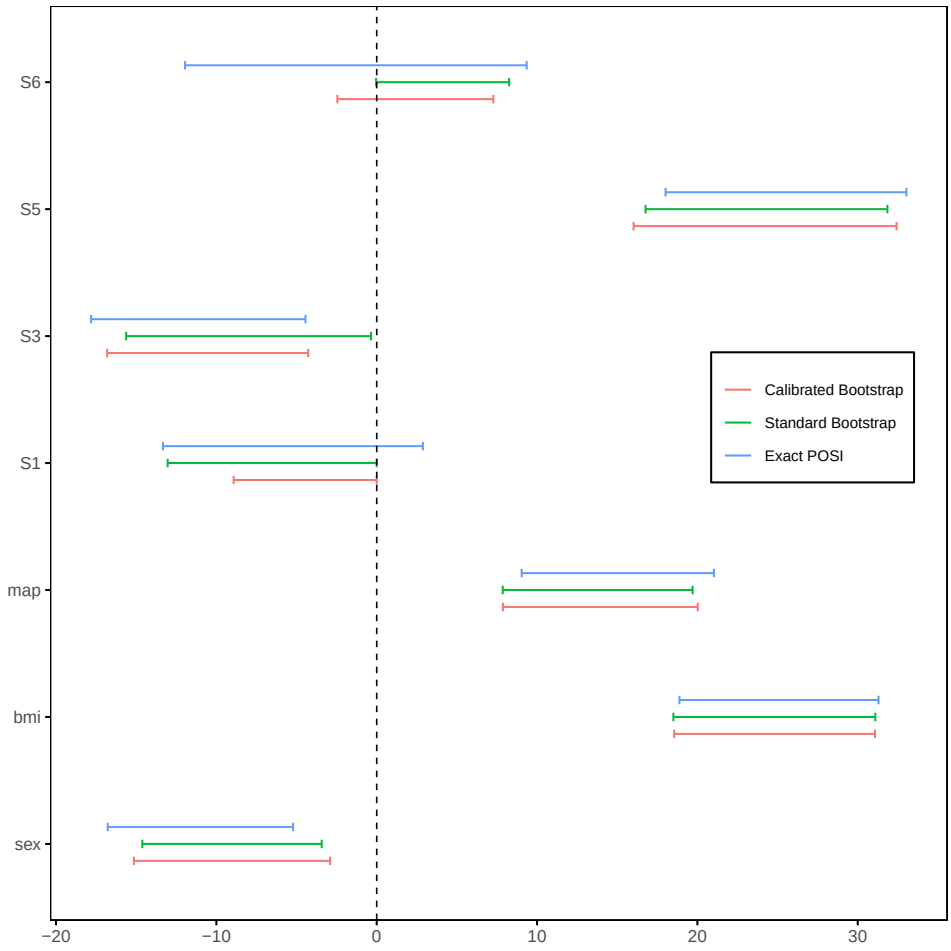


Fig. 5. Constructed 95% confidence intervals for the seven selected variables by Lasso ($\lambda = 520$). The intervals constructed with our proposed CB method are shown in red lines. The intervals constructed with the standard bootstrap are shown in green lines. Exact POSI denotes the post-selection inference approach of Lee et al. (2016), and the constructed intervals are shown in blue lines.

are used for point estimation. Avoiding the limitations associated with conventional bootstrap techniques, the proposed method is shown to achieve valid inference outcomes through calibrated resampling and refinements. To our knowledge, this is the first study to adapt the resampling method for valid inference in finite-sample scenarios. Although the proposed method does involve higher computational costs compared to conventional bootstrap techniques, it remains computationally feasible. Moreover, it can be readily parallelized on modern computer clusters, further enhancing its computational feasibility and efficiency.

The key idea in our development of resampling-based methods for finite-sample valid inference is to find a resampling scheme that guarantees the validity of resulting confidence region for a pre-specified confidence level. In the proposed CB method, we created a stochastic approximation algorithm to find an adaptive m-out-of-n resampling scheme. Alternative resampling schemes and alternative ways of creating samples of bootstrapped estimates can be considered in future research and applications. For example, weighted maximum likelihood estimates with weights drawn from an adaptive Dirichlet($\delta\mathbf{1}$) distribution can be considered, with δ representing the calibration parameter. This method is expected to be especially useful for the case with small observed data samples.

Nevertheless, it is worth noting that obtaining exact marginal inferences for individual parameters can be a complex undertaking, necessitating further research. This highlights the broader challenges of conducting finite-sample valid inference in complex models. In this article, we proposed an approach based on the idea in Martin (2015, 2023b) for "marginalizing out" nuisance parameters. While this method is shown to be valid and empirically efficient, it points to a potential direction that invites creative thoughts on marginal inference, a challenging problem for all existing schools of thought.

The primary focus of this paper is the development of computational methods that facilitate efficient parametric inference. Cella and Martin (2022) developed an IM framework for approximate inference on risk minimizers in a nonparametric context. Given that many modern machine learning applications necessitate inference based on unknown models or loss functions, an intriguing future direction would be to explore whether the insights from this paper can be effectively adapted for efficient inference under the framework proposed by Cella and Martin (2022).

Acknowledgements

The authors are grateful to the editor, the associate editor, and anonymous referees for their insightful, critical, and constructive comments on an earlier version of the paper. Zhang and Jiang are supported in part by U.S. National Institutes of Health grants R01HG010171 and R01MH116527 and National Science Foundation grant DMS-2112711. Liu is supported in part by U.S. National Science Foundation grant DMS-2412629.

References

Abrevaya, J. and Huang, J. (2005) On the bootstrap of the maximum score estimator. *Econometrica*, **73**, 1175–1204.

Arbet, J., McGue, M., Chatterjee, S. and Basu, S. (2017) Resampling-based tests for lasso in genome-wide association studies. *BMC genetics*, **18**, 70–70.

Barndorff-Nielsen, O. E. (1986) Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika*, **73**, 307–322. URL: `http://www.jstor.org/stable/2336207`.

Bickel, P. J. and Freedman, D. A. (1981) Some asymptotic theory for the bootstrap. *The annals of statistics*, **9**, 1196–1217.

— (1983) Bootstrapping regression models with many parameters. *Festschrift for Erich L. Lehmann*, 28–48.

Bickel, P. J., Götze, F. and van Zwet, W. R. (2012) *Resampling fewer than n observations: gains, losses, and remedies for losses*. Springer.

Bickel, P. J. and Sakov, A. (2008) On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica*, 967–985.

Bretagnolle, J. (1983) Lois limites du bootstrap de certaines fonctionnelles. *Ann. Inst. Henri Poincare*, **19**, 281–296.

Cella, L. and Martin, R. (2022) Direct and approximately valid probabilistic inference on a class of statistical functionals. *International Journal of Approximate Reasoning*, **151**, 205–224.

Chakraborty, B., Laber, E. B. and Zhao, Y. (2013) Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics*, **69**, 714–723.

Chatterjee, A. and Lahiri, S. N. (2010) Asymptotic properties of the residualbootstrap for lasso estimators. *Proceedings of the American Mathematical Society*, **138**, 4497–4509. URL: `http://www.jstor.org/stable/41059185`.

— (2011) Bootstrapping lasso estimators. *Journal of the American Statistical Association*, **106**, 608–625. URL: `https://doi.org/10.1198/jasa.2011.tm10159`.

Chernozhukov, V., Chetverikov, D., Kato, K. and Koike, Y. (2023) High-dimensional data bootstrap. *Annual Review of Statistics and Its Application*, **10**, 427–449.

Dempster, A. P. (2008) The dempster–shafer calculus for statisticians. *International Journal of approximate reasoning*, **48**, 365–377.

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

— (1982) *The jackknife, the bootstrap and other resampling plans. CBMS-NSF Regional Conference Series in Applied Mathematics*, vol. 38. SIAM, Philadelphia, PA.

— (2003) Second thoughts on the bootstrap. *Statistical science*, **18**, 135–140.

— (2012) Bayesian inference and the parametric bootstrap. *The annals of applied statistics*, **6**, 1971.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression. *The Annals of Statistics*, **32**, 407–499.

Efron, B. and Tibshirani, R. J. (1993) *An introduction to the bootstrap*. CRC press.

El Karoui, N. and Purdom, E. (2018) Can we trust the bootstrap in high-dimensions? the case of linear models. *The Journal of Machine Learning Research*, **19**, 170–235.

Fisher, R. A. (1973) *Statistical methods and scientific inference, Third Edition*. Hafner Publishing Co.

Freedman, D. A. (1981) Bootstrapping regression models. *The Annals of Statistics*, **9**, 1218–1228. URL: `https://doi.org/10.1214/aos/1176345638`.

Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, **33**, 1–22.

Fu, W. and Knight, K. (2000) Asymptotics for lasso-type estimators. *The Annals of Statistics*, **28**, 1356 – 1378. URL: `https://doi.org/10.1214/aos/1015957397`.

Hannig, J. (2009) On generalized fiducial inference. *Statistica Sinica*, **19**, 491–544.

Hannig, J., Iyer, H., Lai, R. C. and Lee, T. C. (2016) Generalized fiducial inference: A review and new results. *Journal of the American Statistical Association*, **111**, 1346–1361.

Jiang, Y. and Liu, C. (2024) Estimation of over-parameterized models from an auto-modeling perspective. URL: `https://arxiv.org/abs/2206.01824`.

Kosorok, M. R. (2008) *Introduction to empirical processes and semiparametric inference*, vol. 61. Springer.

Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2016) Exact post-selection inference, with application to the lasso. *The Annals of statistics*, **44**, 907–927.

Lehmann, E. L. (1991) *Theory of point estimation*. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, California.

Liu, C. and Martin, R. (2015) Frameworks for prior-free posterior probabilistic inference. *Wiley Interdisciplinary Reviews: Computational Statistics*, **7**, 77–85.

Lopes, M. (2014) A residual bootstrap for high-dimensional regression with near low-rank designs. In *Advances in Neural Information Processing Systems* (eds. Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence and K. Weinberger), vol. 27. Curran Associates, Inc. URL: `https://proceedings.neurips.cc/paper_files/paper/2014/file/7437d136770f5b35194cb46c1653efaa-Paper.pdf`.

Martin, R. (2015) Plausibility functions and exact frequentist inference. *Journal of the American Statistical Association*, **110**, 1552–1561.

— (2021a) An imprecise-probabilistic characterization of frequentist statistical inference. *arXiv preprint arXiv:2112.10904*, **2021**, 1–51.

— (2021b) Inferential models and the decision-theoretic implications of the validity property.

— (2023a) Fiducial inference viewed through a possibility-theoretic inferential model lens. In *ISIPTA'23 proceedings*.

— (2023b) Valid and efficient imprecise-probabilistic inference with partial priors, iii. marginalization. *arXiv preprint arXiv:2309.13454*.

Martin, R. and Liu, C. (2013) Inferential models: A framework for prior-free posterior probabilistic inference. *Journal of the American Statistical Association*, **108**, 301–313.

— (2015) *Inferential models: reasoning with uncertainty*, vol. 145. CRC Press.

Newton, M. A., Polson, N. G. and Xu, J. (2021) Weighted bayesian bootstrap for scalable posterior distributions. *Canadian Journal of Statistics*, **49**, 421–437.

Newton, M. A. and Raftery, A. E. (1994) Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Methodological)*, **56**, 3–26.

Politis, D. N. and Romano, J. P. (1994) Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, 2031–2050.

Reid, S., Tibshirani, R. and Friedman, J. (2016) A study of error variance estimation in lasso regression. *Statistica Sinica*, **26**, 35–67.

Robbins, H. and Monro, S. (1951) A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, **22**, 400 – 407. URL: `https://doi.org/10.1214/aoms/1177729586`.

Rubin, D. B. (1981) The bayesian bootstrap. *The annals of statistics*, 130–134.

— (1987) The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The sir algorithm. *Journal of the American Statistical Association*, **82**, 543–546.

Shafer, G. (1976) *A mathematical theory of evidence*, vol. 42. Princeton university press.

Shao, J. and Tu, D. (2012) *The jackknife and bootstrap*. Springer Science & Business Media.

Singh, K. (1981) On the asymptotic accuracy of efron's bootstrap. *The Annals of Statistics*, 1187–1195.

Singh, K., Xie, M. and Strawderman, W. E. (2007) Confidence distribution (cd): distribution estimator of a parameter. *Lecture Notes-Monograph Series*, 132–150.

Syring, N. and Martin, R. (2019) Calibrating general posterior credible regions. *Biometrika*, **106**, 479–486.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**, 267–288. URL: `http://www.jstor.org/stable/2346178`.

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T. and Posthuma, D. (2021) Genome-wide association studies. *Nature Reviews Methods Primers*, **1**.

van der Vaart, A. W. and Wellner, J. A. (1996) *Weak convergence and empirical processes: with applications to statistics*. Springer-Verlag, New York.

Wellner, J. A. and Zhan, Y. (1996) Bootstrapping z-estimators. *University of Washington Department of Statistics Technical Report*, **308**.

Xie, M.-g. and Singh, K. (2013) Confidence distribution, the frequentist distribution estimator of a parameter: A review. *International Statistical Review*, **81**, 3–39.

Yu, G. and Bien, J. (2019) Estimating the error variance in a high-dimensional linear model. *Biometrika*, **106**, 533–546.

Zabell, S. L. (1992) R. a. fisher and fiducial argument. *Statistical science*, **7**, 369–387.

Zeng, P., Zhao, Y., Qian, C., Zhang, L., Zhang, R., Gou, J., Liu, J., Liu, L. and Chen, F. (2015) Statistical analysis for genome-wide association study. *Journal of Biomedical Research*, **29**, 285–297.