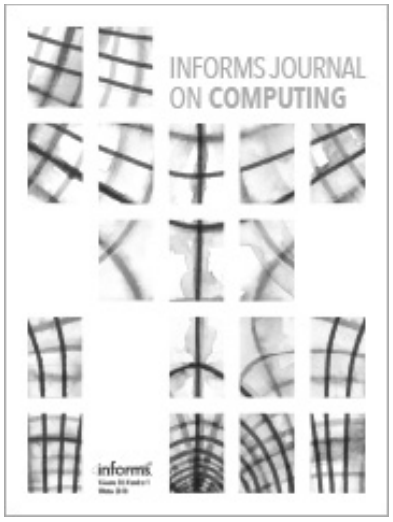


INFORMS Journal on Computing

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

A Shrinkage Approach to Improve Direct Bootstrap Resampling Under Input Uncertainty

Eunhye Song, Henry Lam, Russell R. Barton

To cite this article:
Eunhye Song, Henry Lam, Russell R. Barton (2024) A Shrinkage Approach to Improve Direct Bootstrap Resampling Under Input Uncertainty. INFORMS Journal on Computing

Published online in Articles in Advance 02 Feb 2024

. https://doi.org/10.1287/ijoc.2022.0044

INFORMS JOURNAL ON COMPUTING
Articles in Advance, pp. 1–17
ISSN 1091-9856 (print), ISSN 1526-5528 (online)
https://pubsonline.informs.org/journal/ijoc

A Shrinkage Approach to Improve Direct Bootstrap Resampling Under Input Uncertainty

Eunhye Song,[a,*] Henry Lam,[b] Russell R. Barton[c]

[a] H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332; [b] Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027; [c] Department of Supply Chain and Information Systems, The Pennsylvania State University, University Park, Pennsylvania 16802

*Corresponding author

Contact: esong32@gatech.edu, https://orcid.org/0000-0002-5171-0614 (ES); khl2114@columbia.edu, https://orcid.org/0000-0002-3193-563X (HL); rbarton@psu.edu, https://orcid.org/0000-0002-5054-2006 (RRB)

Received: February 8, 2022
Revised: August 23, 2022; July 16, 2023; November 26, 2023; December 5, 2023
Accepted: December 9, 2023
Published Online in Articles in Advance: February 2, 2024

https://doi.org/10.1287/ijoc.2022.0044

Abstract. Discrete-event simulation models generate random variates from input distributions and compute outputs according to the simulation logic. The input distributions are typically fitted to finite real-world data and thus are subject to estimation errors that can propagate to the simulation outputs: an issue commonly known as input uncertainty (IU). This paper investigates quantifying IU using the output confidence intervals (CIs) computed from bootstrap quantile estimators. The standard direct bootstrap method has overcoverage due to convolution of the simulation error and IU; however, the brute-force way of washing away the former is computationally demanding. We present two new bootstrap methods to enhance direct resampling in both statistical and computational efficiencies using shrinkage strategies to down-scale the variabilities encapsulated in the CIs. Our asymptotic analysis shows how both approaches produce tight CIs accounting for IU under limited input data and simulation effort along with the simulation sample-size requirements relative to the input data size. We demonstrate performances of the shrinkage strategies with several numerical experiments and investigate the conditions under which each method performs well. We also show advantages of nonparametric approaches over parametric bootstrap when the distribution family is misspecified and over metamodel approaches when the dimension of the distribution parameters is high.

History: Accepted by Bruno Tuffin, Area Editor for Simulation.
Funding: This work was supported by the National Science Foundation [CAREER CMMI-1834710, CAREER CMMI-2045400, DMS-1854659, and IIS-1849280].
Supplemental Material: The software that supports the findings of this study is available within the paper and its Supplemental Information (https://pubsonline.informs.org/doi/suppl/10.1287/ijoc.2022.0044) as well as from the IJOC GitHub software repository (https://github.com/INFORMSJoC/2022.0044). The complete IJOC Software and Data Repository is available at https://informsjoc.github.io/.

Keywords: bootstrap resampling • input uncertainty • nonparametric • simulation • shrinkage

1. Introduction

Discrete-event simulation models provide insight on the operational behavior of real systems and forecasts on the behaviors of future systems and policies. Such models typically capture stochastic behaviors using probability distributions fitted to a real-world data sample. In this case, finiteness of the sample introduces errors in the estimated input distributions, affecting the fidelity of the analyses. Barton and Schruben (2001) show the actual coverage of simulation-based confidence intervals (CIs) for the expected waiting time for simple queues can be as low as 20% for nominal 90% CIs with input distributions estimated from samples of size 500. Input uncertainty (IU) analysis characterizes the effect of such input model error on the simulation output variability, in contrast to the simulation error caused by the Monte Carlo randomness incurred in finite-run simulation. A common goal is to develop a CI for mean system performance that reflects both IU and simulation error.

A popular approach to address IU is bootstrapping the input data to approximate the statistical characteristics of real-world samples (Efron 1987). Barton and Schruben (1993) first apply this idea to a discrete-event simulator by using resampled input data to drive simulation runs and estimate the output performance measures and then computing their quantiles to produce a CI. Since then, bootstrapping has been explored in several variants including metamodel-assisted approaches (Barton et al. 2014, Xie et al. 2016) to improve budget allocation (Yi

and Xie 2017) and to estimate the contribution of output variance attributed to IU (Cheng and Holland 1997, Song and Nelson 2015, Lam and Qian 2018).

Quantifying IU via bootstrapping typically involves three steps: (i) resampling (bootstrapping) the input data, (ii) generating simulation outputs at each bootstrap sample and averaging them to compute the bootstrap sample means, and (iii) using the sample means to measure IU by computing relevant statistics. In this paper, we primarily focus on methods where Step (iii) uses the empirical quantiles of the sample means to generate CIs, which encompasses most standard bootstrap methods such as *basic* or *percentile bootstrap* (Efron and Tibshirani 1994). Depending on how Step i is conducted, the methods fall into two categories, nonparameteric versus parametric bootstrap. The former directly samples from the empirical distribution of the input data. The latter makes a parametric assumption about the distribution family of the input data, estimates its parameters, and then draws a bootstrap sample from the fitted distribution. In many cases the nonparametric setting is arguably more desirable as most real-world stochastic phenomena cannot be characterized by parametric families, and parametric assumptions can introduce model errors that can be difficult to characterize. Conversely, the nonparametric case is considered more difficult to analyze as less model assumptions are available to leverage on. For example, in the parametric context, one can readily apply the delta method which involves computing the gradients of the output with respect to the input parameters to construct the CI (Cheng and Holland 1997, 2004; Lin et al. 2015; Morgan et al. 2019). In the nonparametric delta method, the gradients are replaced with the influence function (Hampel 1974) that has an effective dimension growing with the data size, thus adding substantial complexity to its estimation. Another popular parametric approach is to fit a metamodel of the performance measure as a function of input parameters after running replications at some parameter vectors selected in an experiment design (Barton et al. 2014, Xie et al. 2016). However, these approaches run into computational challenges when the dimension of the input parameter vector is high. Typically, it is recommended that the number of parameter vectors in the experiment design should be an order of magnitude larger than its dimension, which may far exceed the number of replications adopted in nonparametric approaches. Moreover, the metamodel fitting process may require optimization, which can be computationally demanding especially when the problem is high dimensional.

Our goal is to devise efficient ways to use nonparametric bootstrap samples to generate tight CIs that address IU. We first elaborate on the methodological challenges that the existing nonparametric bootstrap methods fail to address. The classical quantile-based CIs such as basic or percentile bootstrap CIs can exhibit substantial overcoverage when applied to quantify IU since these techniques are created for a model whose only source of uncertainty is input data, which is equivalent to running infinite replications in Step ii. When the replication size at each bootstrap sample is small relative to the input data size, the bootstrap sample mean is corrupted by simulation error making the empirical quantiles biased and producing a wider CI that overshoots the target coverage probability. To tackle this issue, one can increase the number of replications at each bootstrap sample to wash away the simulation error, but this may add large computational overhead. An alternative approach is to "deconvolute" the simulation and input model errors via density estimation (McIntyre and Stefanski 2011). However, the goal of estimating a full density function is more demanding than generating CIs and these methods typically require kernel selection and additional distributional assumptions on the simulation error.

Motivated by these challenges, this paper investigates nonparametric bootstrap methods that are efficient on two fronts: statistical, namely that our approach generates tight and correct-coverage CIs, and computational, namely that our approach does not require an insurmountable simulation effort to wash away all simulation error. We propose two new methods based on deflating the variability of bootstrap outputs. The first is *sample shrinkage* that systematically down-scales the magnitude of each bootstrap sample mean so that the variability of the resulting outputs only reflects the effect of input model error to generate valid CIs. The second method, *quantile shrinkage*, directly down-scales the upper and lower empirical quantiles of the bootstrap sample mean.

To the best of our knowledge, the shrinkage bootstrap technique is first proposed by Davison and Hinkley (1997) to devise a bootstrap resampling scheme that provides the correct first and second moments for the bootstrap statistic and measurement error, respectively. Under the assumption the measurement error has homoscedastic variance across all bootstrap samples, they derive the expression for the shrinkage factor and its estimator. Flynn and Peters (2004) apply the same idea to analyze clustered medical cost data, where each cluster is regarded as a group. Ng et al. (2013) discusses computational implementation of the shrinkage bootstrap method in Stata. However, none of the reviewed work discusses how the estimation error of the shrinkage factor affects the error in the population statistic the bootstrap method aims to estimate. Doing so requires carefully choosing the bootstrap and simulation sample sizes relative to the input data size. Additionally, using the plug-in estimator for the shrinkage factor makes the resulting shrunk simulation outputs dependent with each other. Yet, none

of the existing work carefully addresses how the dependence affects validity of the bootstrap estimate of the population statistic.

In this paper, we address both sample size requirements and dependence issue in our asymptotic analyses. Specifically, we make the following theoretical contributions to the shrinkage bootstrap literature. First, we provide the requirements for the bootstrap and simulation sample sizes as functions of the input data size so that the resulting shrunk simulation output has the correct marginal distribution that the bootstrap experiment is designed to estimate asymptotically. Second, given our sample-size choices, we show that the dependence among the shrunk simulation output fades away asymptotically as the input data size increases and thus, the proposed shrinkage bootstrap CIs indeed provide the exact asymptotic coverage. Third, our analysis allows the simulation error to be heteroscedastic across the bootstrap samples. Last, we robustify the shrinkage bootstrap CI's empirical performance against the estimation error of the shrinkage factor by computing its lower confidence bound via bootstrapping the same simulation runs made to estimate the shrinkage factor (i.e., no additional simulation cost). Moreover, we provide guidance to determine whether to shrink a bootstrap CI based on the lower bound.

We also report several computational experiments to demonstrate efficiency of the shrinkage CIs including a comparison with metamodel-based IU quantification methods for a large-scale problem (with more than 100 input parameters).

We close this section by contrasting our approach with several recently proposed nonparametric IU quantification methods. First, Song and Nelson (2015) and Lam and Qian (2018) use the bootstrap to estimate the IU variance and construct a normality-based CI. We demonstrate that our CIs not only asymptotically achieve the same half-widths as the normality-based CIs, but also provide tighter empirical CIs for finite sample cases. Second, Glynn and Lam (2018) devise a sectioning approach to construct CIs based on t statistics, which advantageously requires low simulation budget, but could lead to long CIs especially under limited data. Our CIs, conversely, match the normality limit and hence are tighter. Third, Lam and Qian (2016) proposes an optimization approach to build CIs based on the empirical likelihood, which requires algorithmic configurations that are arguably harder to set up than our bootstrap approaches. Fourth, Xie et al. (2021) propose a nonparametric Bayesian approach via Dirichlet mixtures to fit input models, which is different from our frequentist view. Last, Wang et al. (2020) and Song (2021) study metamodels based on nonparametric kernels to incorporate IU into simulation optimization instead of IU quantification addressed here.

The rest of this paper is organized as follows. Section 2 mathematically formulates the IU problem and presents the challenges in direct bootstrap resampling. Section 3 introduces our shrinkage approaches, which are theoretically analyzed in Section 4. We investigate asymptotically efficient simulation sample-size choices for the proposed CIs to achieve tight coverages. Section 5 discusses robustifying the shrinkage CIs. Section 6 presents numerical results followed by concluding remarks in Section 7.

A preliminary conference version of this paper, Barton et al. (2018), explores our idea to enhance direct resampling, however, does not provide theoretical analyses or large-scale experiments. Moreover, the considered CIs are different from those proposed here.

2. Formulation

Suppose the objective of simulation analysis is to estimate a performance measure of the system represented as an expected value of a simulation output. The input cumulative distribution function (cdf) that drives the simulation is denoted by F, which may be multivariate. The simulation output from the rth replication is $Y_r(F) = \psi(F) + \varepsilon_r(F), r = 1, 2, \ldots, R$, where $\psi(F) \triangleq \mathrm{E}[Y_r(F) \mid F]$ and $\varepsilon_r(F)$ represents the simulation error with zero mean and finite variance $\sigma^2(F) \triangleq \mathrm{V}[\varepsilon_r(F) \mid F]$. We take the frequentist view and assume there exists true distribution F^c that generates real-world input data and define $\sigma^2 \triangleq \mathrm{V}[\varepsilon_r(F^c)]$. The input distribution fitted to the data are denoted by $\hat{F}$. IU is introduced when we use $\hat{F}$ in place of F^c to run the simulation and obtain

$$Y_r(\hat{F}) = \psi(\hat{F}) + \varepsilon_r(\hat{F}) = \psi(F^c) + \left(\psi(\hat{F}) - \psi(F^c)\right) + \varepsilon_r(\hat{F}). \tag{1}$$

Note that $\hat{F}$ is a random function constructed from the input data, but F^c is deterministic. Two sources of error are manifested in (1); the first is referred to as input error reflecting that $F^c \neq \hat{F}$ in general; the second is simulation error $\epsilon_r(\hat{F})$.

2.1. Direct Bootstrap Resampling for IU Characterization

To motivate our investigation, we first present the case where ψ can be evaluated exactly without simulation error and then discuss challenges when simulation error is considered. Suppose from F^c, we observe an

independent and identically distributed (i.i.d.) sample of size n. From the input data, the empirical cdf $\hat{F}_0(\cdot) = (1/n)\sum_{i=1}^{n} I(X_i \le \cdot)$ can be defined, where $I(\cdot)$ denotes the indicator function and X_i is the ith observation of the input data. The bootstrap principle stipulates that given $\hat{F}_0$, the quantity $\hat{F}_b$ constructed by resampling n values from the original data with replacement (i.e., $\hat{F}_b(\cdot) = (1/n)\sum_{i=1}^{n} I(X_i^{(b)} \le \cdot)$, where each $X_i^{(b)}$ is sampled from $\hat{F}_0$) satisfies

$$\psi(\hat{F}_b) - \psi(\hat{F}_0) \overset{D}{\approx} \psi(\hat{F}_0) - \psi(F^c), \tag{2}$$

when n is large, where $\overset{D}{\approx}$ denotes approximate equality in distribution, and the left-hand side of (2) is conditional on $\hat{F}_0$. If (2) holds, then we can use the distribution of $\psi(\hat{F}_b) - \psi(\hat{F}_0)$ given $\hat{F}_0$ to approximate that of $\psi(\hat{F}_0) - \psi(F^c)$; notice that the former does not involve F^c. Consequently,

$$P_*(\tau_{\alpha/2} \le \psi(\hat{F}_b) - \psi(\hat{F}_0) \le \tau_{1-\alpha/2}) \approx P(\tau_{\alpha/2} \le \psi(\hat{F}_0) - \psi(F^c) \le \tau_{1-\alpha/2}) = 1-\alpha, \tag{3}$$

where $P_*(\cdot)$ denotes the probability conditional on $\hat{F}_0$, and τ_p is the p-quantile of $\psi(\hat{F}_b) - \psi(\hat{F}_0)$, which approximates the p-quantile of $\psi(\hat{F}_0) - \psi(F^c)$. A $(1-\alpha)$-level CI for $\psi(F^c)$ is approximately

$$[\psi(\hat{F}_0) - \tau_{1-\alpha/2}, \psi(\hat{F}_0) - \tau_{\alpha/2}], \tag{4}$$

where $\tau_{\alpha/2}$ and $\tau_{1-\alpha/2}$ are estimated from the sample quantiles of $\psi(\hat{F}_b) - \psi(\hat{F}_0), b = 1, \ldots, B$, computed via Monte Carlo. We can rewrite (4) as $[2\psi(\hat{F}_0) - q_{1-\alpha/2}, 2\psi(\hat{F}_0) - q_{\alpha/2}]$, where $q_{\alpha/2}$ and $q_{1-\alpha/2}$ are the quantiles of $\psi(\hat{F}_b), b = 1, \ldots, B$. This construction corresponds to the *basic bootstrap* (Davison and Hinkley 1997), which we refer to as BB hereafter.

Alternately, one may take $[q_{\alpha/2}, q_{1-\alpha/2}]$ as a $(1-\alpha)$-CI for $\psi(F^c)$. This scheme can be justified if $\psi(\hat{F}_0) - \psi(\hat{F}^c)$ is symmetrically distributed around zero and (3) holds. Let $\gamma_p(\xi)$ be the p-quantile of a generic random variable ξ, so that $\tau_{\alpha/2}$ and $\tau_{1-\alpha/2}$ defined in (4) can be written as $\gamma_{\alpha/2}(\psi(\hat{F}_b) - \psi(\hat{F}_0))$ and $\gamma_{1-\alpha/2}(\psi(\hat{F}_b) - \psi(\hat{F}_0))$ conditional on $\hat{F}_0$. Then, the lower bound in (4) is

$$\psi(\hat{F}_0) - \gamma_{1-\alpha/2}(\psi(\hat{F}_b) - \psi(\hat{F}_0)). \tag{5}$$

If the distributions of $\psi(\hat{F}_0) - \psi(\hat{F}^c)$ and $\psi(\hat{F}_b) - \psi(\hat{F}_0)$ are symmetric around zero, then $-\gamma_{1-\alpha/2}(\psi(\hat{F}_b) - \psi(\hat{F}_0)) = \gamma_{\alpha/2}(\psi(\hat{F}_b) - \psi(\hat{F}_0))$, and (5) becomes $\psi(\hat{F}_0) + \gamma_{\alpha/2}(\psi(\hat{F}_b) - \psi(\hat{F}_0)) = \gamma_{\alpha/2}(\psi(\hat{F}_b))$. Similarly, the upper bound in (4) becomes $\gamma_{1-\alpha/2}(\psi(\hat{F}_b))$. This procedure is known as the *percentile bootstrap* (Davison and Hinkley 1997). We refer to this method as PB. Notice that PB does not require computing $\psi(\hat{F}_0)$.

The symmetry requirement to justify PB can be relaxed. If there is a monotonically increasing transformation g such that for some v,

$$g(\psi(\hat{F}_b)) - g(\psi(\hat{F}_0)) \overset{D}{\approx} g(\psi(\hat{F}_0)) - g(\psi(F^c)) \overset{D}{\approx} N(0, v^2), \tag{6}$$

then the bootstrap quantiles in the transformed domain are asymptotically valid. The monotonicity of the transform g implies that the same probabilities of coverage apply to the bootstrap sample quantiles. Efron and Tibshirani (1994) illustrate this property for lognormally distributed bootstrap means and claim that PB is more appropriate than BB when $\psi(\hat{F}_0)$ and $\psi(\hat{F}_b)$ given $\hat{F}_0$ have asymmetric distributions and g satisfying (6) exists. It is not necessary to identify g; its existence suffices. We observe this characteristic in our computational study in Section 6.

Because PB CIs use the empirical bootstrap quantiles directly, they avoid the risk of CIs falling outside the support of $\psi(\hat{F}_0)$. For example, if $\psi(F) > 0$ for any F, then the positive skewness of $\psi(\hat{F}_b)$ can make the lower confidence bound of BB negative, but not for PB.

Unfortunately, $\psi(\cdot)$ cannot be computed exactly in a stochastic simulation setting and can only be estimated by $\overline{Y}_R(F) \triangleq \sum_{r=1}^{R} Y_r(F)/R$ given $Y_1(F), Y_2(F), \ldots, Y_R(F)$. Therefore, instead of (2), we hope when n is large,

$$\overline{Y}_R(\hat{F}_b) - \overline{Y}_{R_1}(\hat{F}_0) \overset{D}{\approx} \overline{Y}_{R_0}(\hat{F}_0) - \psi(F^c) \tag{7}$$

conditional on $\hat{F}_0$ in the left-hand side. Note that R_0 is the simulation replication size for the point estimator of $\psi(\hat{F}_0)$ in the right-hand side of (7), R is the replication size for each bootstrap sample, and R_1 is the replication size for the estimate of $\psi(\hat{F}_0)$ in the bootstrap distribution. The choices of R_0, R, and R_1 can be different, which are discussed in detail in subsequent sections. If (7) holds, then we can use the bootstrap principle similar to (3). Namely, we take the $\alpha/2$ and $1-\alpha/2$ sample quantiles of $\overline{Y}_R(\hat{F}_b) - \overline{Y}_{R_1}(\hat{F}_0), b = 1, \ldots, B$, denoted by $\hat{\tau}_{\alpha/2}$ and

$\hat{\tau}_{1-\alpha/2}$, respectively, and construct a $(1-\alpha)$-level BB CI for $\psi(F^c)$ as

$$[\overline{Y}_{R_0}(\hat{F}_0) - \hat{\tau}_{1-\alpha/2}, \overline{Y}_{R_0}(\hat{F}_0) - \hat{\tau}_{\alpha/2}]. \tag{8}$$

Similarly, the PB CI can be constructed by taking the $\alpha/2$ and $1-\alpha/2$ sample quantiles of $\overline{Y}_R(\hat{F}_1), \ldots, \overline{Y}_R(\hat{F}_B)$, denoted by $\hat{q}_{\alpha/2}$ and $\hat{q}_{1-\alpha/2}$, respectively, and returning

$$[\hat{q}_{\alpha/2}, \hat{q}_{1-\alpha/2}]. \tag{9}$$

As observed earlier, we do not need to estimate $\overline{Y}_{R_0}(\hat{F}_0)$ or $\overline{Y}_{R_1}(\hat{F}_0)$ for (9), which can be quite beneficial when R_0 and R_1 turn out to be large.

2.2. Inefficiency of Direct Bootstrap Resampling

The key challenge of direct resampling is to characterize the requirements for R, R_0, and R_1 such that (7) holds asymptotically as $n \to \infty$. To tackle this, it is helpful to first understand the asymptotic normality of a point estimator $\overline{Y}_{R_0}(\hat{F}_0)$, which is subject to both input and simulation errors. Using the delta method, it can be shown that $\overline{Y}_{R_0}(\hat{F}_0)$ satisfies a central limit theorem (CLT): $\overline{Y}_{R_0}(\hat{F}_0) \overset{D}{\approx} N(\psi(F^c), V^2 + \sigma^2/R_0)$ as n and R_0 get large, where $V^2 \triangleq \mathrm{V}(\psi(\hat{F}_0))$ (Cheng and Holland 1997). Thus,

$$\left[\overline{Y}_{R_0}(\hat{F}_0) - z_{1-\alpha/2}\sqrt{V^2 + \frac{\sigma^2}{R_0}}, \overline{Y}_{R_0}(\hat{F}_0) + z_{1-\alpha/2}\sqrt{V^2 + \frac{\sigma^2}{R_0}}\right] \tag{10}$$

is a valid $(1-\alpha)$-level CI, where z_α is the α-quantile of the standard normal distribution. Here, V^2 typically scales reciprocally with the input sample size (Song et al. 2014, Lam and Qian 2018). Cheng and Holland (1997) suggest estimating V^2 (and σ^2) via bootstrap, but they focus on parametric input models.

There are challenges when directly using (10). Because estimating V^2 via bootstrap involves resampling the input data and running simulations for each resample, the computational load can be large. Moreover, even when this computational effort is reduced by, for instance, suitably subsampling the input data (Lam and Qian 2018), the question remains on the best allocation of the simulation effort. A delta method–based CI can run into undercoverage issues with finite samples due to the inadequacy of the linear approximation of the performance measure. This motivates devising quantile-based bootstrap methods that can have better finite-sample performance.

Given input sample size n, the width of (10) is minimized when R_0 is large enough to wash away the simulation error. In this case, the CI becomes

$$[\overline{Y}_{R_0}(\hat{F}_0) - z_{1-\alpha/2}V, \overline{Y}_{R_0}(\hat{F}_0) + z_{1-\alpha/2}V], \tag{11}$$

which has a half-width $z_{1-\alpha/2}V$. Suppose now that we adopt BB, with the same point estimator, $\overline{Y}_{R_0}(\hat{F}_0)$, in (8). The bootstrap principle stipulates that the resampled $\overline{Y}_R(\hat{F}_b) - \overline{Y}_{R_1}(\hat{F}_0)$ in (7) should have a distribution mimicking $\overline{Y}_{R_0}(\hat{F}_0) - \psi(F^c)$, which implies that R should be chosen as R_0 (and R_1 chosen as another large number enough to wash away the simulation error in $\overline{Y}_{R_1}(\hat{F}_0)$). This approach would then impose a heavy computational burden as we need to run $R = R_0$ new simulation replications for each of the B bootstrap resamples, which amounts to $(B+1)R_0$ replications in total.

Suppose that we are less ambitious and are content with a half-width approximately $z_{1-\alpha/2}\sqrt{V^2 + \sigma^2/R}$, with R relatively small. In this case, we can use a point estimator $\overline{Y}_R(\hat{F}_0)$ (i.e., $R_0 = R$), and resample estimators $\overline{Y}_R(\hat{F}_b)$ in (7). In the case of BB, the total budget becomes $BR + R_1$, and we need R_1 to be large enough so that the bootstrap distribution of $\overline{Y}_R(\hat{F}_b) - \overline{Y}_{R_1}(\hat{F}_0)$ approximates the distribution of $\overline{Y}_R(\hat{F}_0) - \psi(F^c)$ sufficiently well in the limit. Conversely, if we have computed $\overline{Y}_{R_1}(\hat{F}_0)$ using a large R_1 in the bootstrap distribution, it is wasteful not to use it as a point estimator in constructing a CI. In the case of PB, we can avoid computing $\overline{Y}_{R_1}(\hat{F}_0)$ and hence the simulation budget is BR. Nonetheless, in either approach, one may wonder whether we can minimize the half-width down to $z_{1-\alpha/2}V$ by leveraging the structure of the bootstrapped simulation, instead of $z_{1-\alpha/2}\sqrt{V^2 + \sigma^2/R}$.

Based on the previous discussions, our main investigation is to design schemes that can both generate a CI with the minimum half-width, while controlling R to be small. In the remainder of the paper, we take $R_0 = R_1$ and reuse the estimates $\overline{Y}_{R_0}(\hat{F}_0) = \overline{Y}_{R_1}(\hat{F}_0)$ in the two sides of (7) and choose a relatively small R to construct a CI with a half-width that matches the approximate minimum length $z_{1-\alpha/2}V$ dictated by the CLT. Such schemes would be efficient both statistically (short half-widths) and computationally (a small simulation budget). Designing these schemes is the focus of the next section.

3. Shrinkage Bootstrap to Strengthen Direct Resampling

We propose two shrinkage approaches that aim to properly deflate the variability of the bootstrapped outputs to match the minimal CI half-width implied by the CLT. These approaches are paired with either BB or PB, yielding four candidate methods. Proofs of the theoretical results in this section are included in Section OS.1 of the online supplement.

3.1. Sample Shrinkage Bootstrap

Our first approach is the *sample shrinkage* procedure that removes the excess variation caused by simulation error in each $\overline{Y}_R(\hat{F}_b)$. Namely, our idea is to shrink each $\overline{Y}_R(\hat{F}_b)$, $b = 1, \ldots, B$, toward the grand mean, $\overline{\overline{Y}}_R = \frac{1}{B}\sum_{b=1}^{B} \overline{Y}_R(\hat{F}_b)$, as

$$\hat{Y}_R(\hat{F}_b) = c\overline{\overline{Y}}_R + (1-c)\overline{Y}_R(\hat{F}_b). \tag{12}$$

The quantity, c, is the *shrinkage factor* that satisfies

$$(1-c)^2 \triangleq \frac{B}{B-1}\frac{\mathrm{V}_*(\psi(\hat{F}_b))}{\mathrm{V}_*(\psi(\hat{F}_b)) + \mathrm{E}_*[\sigma^2(\hat{F}_b)]/R}, \tag{13}$$

where $\mathrm{E}_*[\cdot]$ and $\mathrm{V}_*(\cdot)$ denote the expectation and variance over $\hat{F}_b$ conditional on $\hat{F}_0$ or equivalently the original data, that is, $\mathrm{E}_*[\cdot] = \mathrm{E}[\cdot \mid \hat{F}_0]$ and $\mathrm{V}_*(\cdot) = \mathrm{V}(\cdot \mid \hat{F}_0)$. In general, $\mathrm{V}_*(\psi(\hat{F}_b))$ and $\mathrm{E}_*[\sigma^2(\hat{F}_b)]$ are unknown. Thus, we plug in their estimates to define

$$1 - \hat{c} \triangleq \sqrt{\max\left\{0, \frac{B}{B-1} - \frac{SS_{within}}{R(R-1)SS_{between}}\right\}}, \tag{14}$$

where $SS_{within} \triangleq \sum_{b=1}^{B}\sum_{r=1}^{R}(Y_r(\hat{F}_b) - \overline{Y}_R(\hat{F}_b))^2$ and $SS_{between} \triangleq \sum_{b=1}^{B}(\overline{Y}_R(\hat{F}_b) - \overline{\overline{Y}}_R)^2$ are the within-group and between-group sums of squares of the bootstrapped simulation outputs, respectively. Then, a $1-\alpha$ *sample shrinkage basic bootstrap* (SSB) CI is constructed as $[\overline{Y}_{R_0}(\hat{F}_0) - \tilde{\tau}_{1-\alpha/2}, \overline{Y}_{R_0}(\hat{F}_0) - \tilde{\tau}_{\alpha/2}]$, where $\tilde{\tau}_p$ is the sample p-quantile of $\hat{Y}_R(\hat{F}_b) - \overline{Y}_{R_0}(\hat{F}_0), b = 1, \ldots, B$. In other words, the quantiles, $\hat{\tau}_{\alpha/2}$ and $\hat{\tau}_{1-\alpha/2}$, of $\overline{Y}_R(\hat{F}_b) - \overline{Y}_{R_0}(\hat{F}_0), b = 1, \ldots, B$, in (8) are replaced with their shrunk counterparts.

We first establish that the variance of the shrunk resampled simulation outputs, $\mathrm{V}_*(\hat{Y}_R(\hat{F}_b))$, matches the variance contributed from IU only, namely $\mathrm{V}_*(\psi(\hat{F}_b))$.

Proposition 1 (Shrinkage Factor Matching). *Given* $\hat{F}_0$, *suppose we modify* (13) *to*

$$(1-c)^2 = \frac{B}{B-1}\frac{\mathrm{V}_*(\psi(\hat{F}_b))}{\mathrm{V}_*(\psi(\hat{F}_b)) + \mathrm{E}_*[\sigma^2(\hat{F}_b)]/R} - \frac{1}{B-1}. \tag{15}$$

Then, $\mathrm{V}_*(\hat{Y}_R(\hat{F}_b)) = \mathrm{V}_*(\psi(\hat{F}_b))$.

Note that (15) is the same as the right-hand side of (13) except for the last term, $-1/(B-1)$, which is negligible for large B. To estimate (15), we adopt $SS_{within}/(B(R-1))$ and $SS_{between}/(B-1) - SS_{within}/(BR(R-1))$ as unbiased estimators of $\mathrm{E}_*[\sigma^2(\hat{F}_b)]$ and $\mathrm{V}_*(\psi(\hat{F}_b))$, respectively. Proposition 1 does not require homoscedasticity of the simulation errors, that is, $\sigma^2(\hat{F}_b)$ needs not be identical for all b.

The same idea applies to constructing a $(1-\alpha)$-level *sample shrinkage percentile bootstrap* (SSP) CI. Starting from the shrunk resampled simulation outputs $\hat{Y}_R(\hat{F}_b)$, $b = 1, \ldots, B$, we obtain their $\alpha/2$ and $1-\alpha/2$ sample quantiles $\tilde{q}_{\alpha/2}$ and $\tilde{q}_{1-\alpha/2}$ and return $[\tilde{q}_{\alpha/2}, \tilde{q}_{1-\alpha/2}]$. In either SSB or SSP, computing $\hat{c}$ and constructing the sample shrinkage interval requires no additional simulation runs compared with BB or PB. Similar to PB, SSP does not require the point estimate of $\psi(\hat{F}_0)$ saving R_0 replications at $\hat{F}_0$.

Algorithm 1 (Sample Shrinkage Bootstrap CIs (SSB and SSP))

1: [**Point estimator**] Run R_0 replications of the simulator using $\hat{F}_0$ as the input model to obtain $\overline{Y}_{R_0}(\hat{F}_0) = \sum_{r=1}^{R_0} Y_r(\hat{F}_0)/R_0$.
2: **for** $b = 1, 2, \ldots, B$ **do**
3: Generate $\hat{F}_b$ by resampling $\hat{F}_0$ n times.
4: Using $\hat{F}_b$, generate $Y_1(\hat{F}_b), Y_2(\hat{F}_b), \ldots, Y_R(\hat{F}_b)$ and compute $\overline{Y}_R(\hat{F}_b) = \sum_{r=1}^{R} Y_r(\hat{F}_b)/R$.
5: Using $Y_r(\hat{F}_b), b = 1, 2 \ldots, B, r = 1, 2, \ldots, R$, calculate $\hat{c}$ from (14).
6: **for** $b = 1, \ldots, B$, compute $\hat{Y}_R(\hat{F}_b) = \hat{c}\overline{\overline{Y}}_R + (1-\hat{c})\overline{Y}_R(\hat{F}_b)$, where $\overline{\overline{Y}} = \sum_{b=1}^{B} \overline{Y}_R(\hat{F}_b)/B$.
7: **CI construction**:

8: [**Sample shrinkage basic bootstrap (SSB) CI**] Find the empirical $\alpha/2$ and $1-\alpha/2$ quantiles of $\hat{Y}_R(\hat{F}_1)-\overline{Y}_{R_0}(\hat{F}_0)$, $\hat{Y}_R(\hat{F}_2)-\overline{Y}_{R_0}(\hat{F}_0),\ldots,\hat{Y}_R(\hat{F}_B)-\overline{Y}_{R_0}(\hat{F}_0)$, denoted by $\tilde{\tau}_{\alpha/2}$ and $\tilde{\tau}_{1-\alpha/2}$, respectively, and return $[\overline{Y}_{R_0}(\hat{F}_0)-\tilde{\tau}_{1-\alpha/2},\overline{Y}_{R_0}(\hat{F}_0)-\tilde{\tau}_{\alpha/2}]$.
9: [**Sample shrinkage percentile bootstrap (SSP) CI**] Find the empirical $\alpha/2$ and $1-\alpha/2$ quantiles of $\hat{Y}_R(\hat{F}_1)$, $\hat{Y}_R(\hat{F}_2),\ldots,\hat{Y}_R(\hat{F}_B)$, denoted by $\tilde{q}_{\alpha/2}$ and $\tilde{q}_{1-\alpha/2}$, respectively, and return $[\tilde{q}_{\alpha/2},\tilde{q}_{1-\alpha/2}]$.

Algorithm 1 summarizes how to compute the sample shrinkage bootstrap CIs.

3.2. Quantile Shrinkage Bootstrap

Our second proposal is *quantile shrinkage* bootstrap that directly shrinks the involved quantiles of the bootstrap sample means. Recall that the BB CI introduced in Section 2.1 is $[\overline{Y}_{R_0}(\hat{F}_0)-\hat{\tau}_{1-\alpha/2},\overline{Y}_{R_0}(\hat{F}_0)-\hat{\tau}_{\alpha/2}]$, where $\hat{\tau}_{\alpha/2}$ and $\hat{\tau}_{1-\alpha/2}$ are the $\alpha/2$ and $1-\alpha/2$ sample quantiles of $\overline{Y}_R(\hat{F}_b)-\overline{Y}_{R_0}(\hat{F}_0), b=1,\ldots,B$. We propose to shrink $\hat{\tau}_{\alpha/2}$ and $\hat{\tau}_{1-\alpha/2}$ to $(1-c)\hat{\tau}_{\alpha/2}$ and $(1-c)\hat{\tau}_{1-\alpha/2}$, respectively, and return

$$[\overline{Y}_{R_0}(\hat{F}_0)-(1-c)\hat{\tau}_{1-\alpha/2},\overline{Y}_{R_0}(\hat{F}_0)-(1-c)\hat{\tau}_{\alpha/2}] \tag{16}$$

as the $(1-\alpha)$-level *quantile shrinkage basic bootstrap* (QSB) CI for $\psi(F^c)$.

For the PB version, we take $\hat{q}_{\alpha/2}$ and $\hat{q}_{1-\alpha/2}$, the $\alpha/2$ and $1-\alpha/2$ quantiles of $\overline{Y}_R(\hat{F}_b), b=1,\ldots,B$, and shrink them to $(1-c)\hat{q}_{\alpha/2}$ and $(1-c)\hat{q}_{1-\alpha/2}$. Then we output the $(1-\alpha)$-level *quantile shrinkage percentile bootstrap* (QSP) CI for $\psi(F^c)$ as

$$[c\overline{Y}_{R_0}(\hat{F}_0)+(1-c)\hat{q}_{\alpha/2},c\overline{Y}_{R_0}(\hat{F}_0)+(1-c)\hat{q}_{1-\alpha/2}]. \tag{17}$$

Similar to PB, (17) is reasoned from the distributional symmetry of $\overline{Y}_R(\hat{F}_0)-\psi(\hat{F}^c)$ and $\overline{Y}_R(\hat{F}_b)-\overline{Y}_{R_0}(\hat{F}_0)$, so that $-\tau_{1-\alpha/2}(\overline{Y}_R(\hat{F}_b)-\overline{Y}_{R_0}(\hat{F}_0))=\tau_{\alpha/2}(\overline{Y}_R(\hat{F}_b)-\overline{Y}_{R_0}(\hat{F}_0))$. This translates (16) to $[\overline{Y}_{R_0}(\hat{F}_0)+(1-c)\hat{\tau}_{\alpha/2},\overline{Y}_{R_0}(\hat{F}_0)+(1-c)\hat{\tau}_{1-\alpha/2}]$, which is equivalent to (17). Observe that each end point of (17) is a weighted average of $\overline{Y}_{R_0}(\hat{F}_0)$ and $\hat{q}_{\alpha/2}$ or $\hat{q}_{1-\alpha/2}$, where the weight given to the former is precisely the shrinkage factor, c. Once again, we estimate c by $\hat{c}$ in (14). Compared with SP discussed in Section 3.1, the point estimator, $\overline{Y}_{R_0}(\hat{F}_0)$, is not canceled out in (17). In other words, the benefit of not requiring calculation of the point estimate vanishes in this construction.

Algorithm 2 details the computation of the two quantile shrinkage bootstrap CIs.

Algorithm 2 (Quantile Shrinkage Bootstrap CIs (QSB and QSP))
1: Run Steps 1–5 of Algorithm 1.
2: **CI construction**:
3: [**Quantile shrinkage basic bootstrap (QSB) CI**] Find the empirical $\alpha/2$ and $1-\alpha/2$ quantiles of $\overline{Y}_R(\hat{F}_b)-\overline{Y}_{R_0}(\hat{F}_0)$, denoted by $\hat{\tau}_{\alpha/2}$ and $\hat{\tau}_{1-\alpha/2}$ respectively, and return $[\overline{Y}_{R_0}(\hat{F}_0)-(1-\hat{c})\hat{\tau}_{1-\alpha/2},\overline{Y}_{R_0}(\hat{F}_0)-(1-\hat{c})\hat{\tau}_{\alpha/2}]$.
4: [**Quantile shrinkage percentile bootstrap (QSP) CI**] Find the empirical $\alpha/2$ and $1-\alpha/2$ quantiles of $\overline{Y}_R(\hat{F}_b)$, denoted by $\hat{q}_{\alpha/2}$ and $\hat{q}_{1-\alpha/2}$ respectively, and return $[\hat{c}\overline{Y}_{R_0}(\hat{F}_0)+(1-\hat{c})\hat{q}_{\alpha/2},\hat{c}\overline{Y}_{R_0}(\hat{F}_0)+(1-\hat{c})\hat{q}_{1-\alpha/2}]$.

4. Theory for Half-Width Minimization and Overcoverage Avoidance

This section establishes the validity of the shrinkage bootstrap CIs in Section 3. In particular, we show that the obtained intervals have half-widths that are asymptotically equivalent to the one stipulated by $N(0,V^2)$. We focus on the basic bootstrap versions (SSB and QSB) throughout our analyses in this section, noting that similar analyses can be used for the PB versions. All proofs of theoretical results in this section are presented in Section OS.1 of the online supplement.

We adopt $\Rightarrow$ to denote convergence in distribution, and $\xrightarrow{p}$ for convergence in probability. For random variables A and B, we say $A\mid\hat{F}_0\Rightarrow B$ in probability, if

$$P_*(A\le x)\xrightarrow{p}P(B\le x)\ \text{ for any }\ x\in\mathbb{R},\ \text{where}\ \ P_*(\cdot)\triangleq P(\cdot\mid\hat{F}_0). \tag{18}$$

Note that $P_*(A\le x)$ in (18) is random as the probability is conditional on $\hat{F}_0$ and thus invokes the convergence in probability. See lemma 10.11 in Kosorok (2007) for the justification of (18) to define the notion of conditional weak convergence in probability. Additionally, for any two positive sequences $\{a_k\}$ and $\{b_k\}$ indexed by k, we say $a_k=o(b_k)$ if $a_k/b_k\to 0$ as $k\to\infty$, $a_k=\omega(b_k)$ if $a_k/b_k\to\infty$ as $k\to\infty$, and $a_k=\Theta(b_k)$ if $\underline{C}\le a_k/b_k\le\overline{C}$ for some $0<\underline{C}\le\overline{C}<\infty$ and large enough k. For random sequences $\{a_k\}$ and deterministic sequence $\{b_k\}$, we say $a_k=o_p(b_k)$ if $a_k/b_k\xrightarrow{p}0$, and $a_k=O_p(b_k)$ if for any $\epsilon>0$, there exist $M,N>0$ such that $P(|a_k/b_k|\le M)>1-\epsilon$ for $k>N$.

Here we state some general assumptions on ψ to establish the theory.

Assumption 1 (Central Limit Theorem). *The performance measure,* $\psi(\cdot)$*, satisfies* $\sqrt{n}(\psi(\hat{F}_0) - \psi(F^c)) \Rightarrow N(0, \sigma_I^2)$ *as* $n \to \infty$*, where* $\sigma_I^2 > 0$.

Assumption 2 (Bootstrap Principle). *We have* $\sqrt{n}(\psi(\hat{F}_b) - \psi(\hat{F}_0)) \mid \hat{F}_0 \Rightarrow N(0, \sigma_I^2)$ *in probability as* $n \to \infty$.

Assumption 1 is a mild condition satisfied by most performance measures and forms the basis to justify the CLT-based CI. Assumption 2 is the bootstrap principle, which states that replacing F^c and $\hat{F}_0$ with $\hat{F}_0$ and $\hat{F}_b$, respectively, satisfies the same asymptotic limit. Both Assumptions 1 and 2 are implied by Hadamard differentiability of $\psi(\cdot)$ that is standard in the bootstrap literature (see theorems 20.8 and 23.9 in Van der Vaart (2000)).

The next assumption is regarding the behavior of the bootstrap.

Assumption 3 (Bootstrap Principle for Input and Stochastic Uncertainties). *We have* $n\mathrm{V}(\psi(\hat{F}_0)) \to \sigma_I^2$ *and* $\mathrm{E}[\sigma^2(\hat{F}_0)] \to \sigma^2$ *as* $n \to \infty$*. Moreover, we have* $n\mathrm{V}_*(\psi(\hat{F}_b)) \xrightarrow{p} \sigma_I^2$, $\mathrm{E}_*[\sigma^2(\hat{F}_b)] \xrightarrow{p} \sigma^2$, *and* $\mathrm{E}_*[\psi(\hat{F}_b)] - \psi(\hat{F}_0) = o_p(1/\sqrt{n})$ *as* $n \to \infty$.

Assumption 3 implies that the bootstrap variance, $\mathrm{V}_*(\psi(\hat{F}_b))$, closely approximates IU variance $\mathrm{V}(\psi(\hat{F}_0))$, a consequence of Assumptions 1 and 2 with further integrability conditions. Assumption 3 also states that both $\mathrm{E}[\sigma^2(\hat{F}_0)]$ and $\mathrm{E}_*[\sigma^2(\hat{F}_b)]$ approach $\sigma^2 = \mathrm{V}(\varepsilon_r(F^c))$, which is reasoned from the closeness of $\hat{F}_b$ and $\hat{F}_0$ to F^c. Finally, $\mathrm{E}_*[\psi(\hat{F}_b)] - \psi(\hat{F}_0) = o_p(1/\sqrt{n})$ follows from the conditional CLT in Assumption 2, so that the conditional bias, $\mathrm{E}_*[\psi(\hat{F}_b)] - \psi(\hat{F}_0)$, should be of smaller order than the CLT scaling, $1/\sqrt{n}$.

Recall from Section 2 that we aim to choose small R, whereas R_0 may be a relatively large number; we make these choices precise here. By a small number we mean $R = \Theta(n)$, that is, comparable to the input sample size, and by a large number we mean $R_0 = \omega(n)$, that is, of a larger order than the input sample size. In particular, we assume there is constant $p > 0$ such that $R/n \to p$. Under this setting, we have the following result.

Proposition 2 (Shrinkage Factor Representation). *Suppose Assumption* 3 *holds and* $R/n \to p$ *for some fixed constant* $p > 0$*. Let* $1 - c' \triangleq \sqrt{\frac{\mathrm{V}_*(\overline{Y}_R(\hat{F}_b)) - \mathrm{E}_*[\sigma^2(\hat{F}_b)]/R}{\mathrm{V}_*(\overline{Y}_R(\hat{F}_b))}}$*, then we have* $1 - c' = \sqrt{\frac{\sigma_I^2/n}{\sigma_I^2/n + \sigma^2/R}}(1 + o_p(1))$.

Note that c' differs from c in (13) only by a factor of $\sqrt{\frac{B}{B-1}}$, which is negligible for moderate B. Additional moment conditions follow next; we let $v^3(F) \triangleq \mathrm{E}[\varepsilon^3(F) \mid F]$ be the third conditional moment of the simulation error given F, and $v^3 \triangleq v^3(F^c)$ for simplicity.

Assumption 4 (Moment Conditions). *We assume* $\mathrm{E}_*[(\psi(\hat{F}_b) - \mathrm{E}_*\psi(\hat{F}_b))^4] = O_p(1/n^2)$, $\mathrm{E}_*[(Y_r(\hat{F}_b) - \psi(\hat{F}_b))^4] = O_p(1)$ *and* $\mathrm{E}_*[\sigma^4(\hat{F}_b)] = O_p(1)$ *as* $n \to \infty$*. Moreover,* $\sigma^2(\hat{F}_b) \xrightarrow{p} \sigma^2$ *and* $v^3(\hat{F}_b) \xrightarrow{p} v^3$ *as* $n \to \infty$.

The scaling, $\mathrm{E}_*[(\psi(\hat{F}_b) - \mathrm{E}_*\psi(\hat{F}_b))^4] = O_p(1/n^2)$, is reasoned from the conditional CLT in Assumption 2 so that $(\psi(\hat{F}_b) - \mathrm{E}_*\psi(\hat{F}_b))^4$ is of order $1/n^2$ while $O_p(1)$ for $\mathrm{E}_*[(Y_r(\hat{F}_b) - \psi(\hat{F}_b))^4]$ and $\mathrm{E}_*[\sigma^4(\hat{F}_b)]$ follow from a finite simulation variance together with suitable integrability conditions. The convergences of $\sigma^2(\hat{F}_b)$ and $v^3(\hat{F}_b)$ to σ^2 and v^3 are reasoned from the closeness of $\hat{F}_b$ to F^c as for Assumption 3. These convergences implicitly involve $\hat{F}_0$. The following result shows that $\hat{c}$ defined in (14) has the same limit in probability as c in (13). Note that (14) can be rewritten as $1 - \hat{c} = \sqrt{\max\{0, (B(\hat{V} - \hat{W}/R))/((B-1)\hat{V})\}}$, where $\hat{V} \triangleq SS_{between}/(B-1)$ and $\hat{W} \triangleq SS_{within}/(B(R-1))$.

Proposition 3 (Estimating Shrinkage Factor). *Suppose Assumptions* 3 *and* 4 *hold,* $R/n \to p$ *for some fixed constant* $p > 0$*, and* $B = \omega(1)$*. Then, we have*

$$\hat{V} - \frac{\hat{W}}{R} = \left(\mathrm{V}_*(\overline{Y}_R(\hat{F}_b)) - \frac{\mathrm{E}_*[\sigma^2(\hat{F}_b)]}{R}\right)(1 + o_p(1)),\ \hat{V} = \mathrm{V}_*(\overline{Y}_R(\hat{F}_b))(1 + o_p(1)), \tag{19}$$

$$\textit{and } 1 - \hat{c} = (1 - c)(1 + o_p(1)) = \sqrt{\frac{\sigma_I^2/n}{\sigma_I^2/n + \sigma^2/R}}(1 + o_p(1)). \tag{20}$$

Last, we make an additional assumption regarding nondegeneracy of simulation error.

Assumption 5 (Nondegenerate Simulation Variance). *We have* $\sigma^2 > 0$.

Although Assumption 5 only imposes a positivity condition on the simulation variance at F^c, combined with Assumption 4, it implies that such a condition also holds for $\sigma^2(\hat{F}_b)$ with an overwhelming probability as n increases.

The last assumption is on the choice of R and R_0 relative to n.

Assumption 6 (Simulation Sample Size Choices). *We choose R and R_0 such that $R/n \to p$ for some fixed constant $p > 0$, $R_0 = \omega(n)$, and $B = \omega(1)$.*

We are now ready to state one of the main results of this section.

Theorem 1 (Asymptotic Limit and Minimal Variance of Shrinkage Bootstrap). *Suppose Assumptions 2–6 hold. Then, $\sqrt{n}((1-\hat{c})\overline{Y}_R(\hat{F}_b) + \hat{c}\overline{\overline{Y}}_R - \overline{Y}_{R_0}(\hat{F}_0)) \mid \hat{F}_0 \Rightarrow N(0,\sigma_I^2)$ in probability.*

Theorem 1 stipulates that $\sqrt{n}((1-\hat{c})\overline{Y}_R(\hat{F}_b) + \hat{c}\overline{\overline{Y}}_R - \overline{Y}_{R_0}(\hat{F}_0))$ converges to $N(0,\sigma_I^2)$ in the asymptotic limit, which matches the variability of the minimal-variance Gaussian variable that induces the normal CI in (11). To translate this into validity of our bootstrap schemes, we need to show that the empirical α-quantile of $\{\hat{Y}(\hat{F}_b) - \overline{Y}_{R_0}(\hat{F}_0)\}_{b=1}^B$ denoted by $\tilde{\tau}_\alpha$ converges to the corresponding quantile of $N(0,\sigma_I^2)$. This would conclude our bootstrap interval matches the normality-based interval with the minimal variance, and moreover establish the asymptotically exact coverage of our shrinkage-based bootstrap CI methods.

Because $\{\sqrt{n}(\hat{Y}(\hat{F}_b) - \overline{Y}_{R_0}(\hat{F}_0))\}_{b=1}^B$ are dependent due to our introduction of $\hat{c}$ (even when conditional on $\hat{F}_0$), we need to ensure that their empirical cdf converges to $N(0,\sigma_I^2)$ despite the dependence. Here, we develop a uniform convergence result for the empirical cdf constructed from $\{\sqrt{n}(\hat{Y}(\hat{F}_b) - \overline{Y}_{R_0}(\hat{F}_0))\}_{b=1}^B$ given $\hat{F}_0$ denoted by $\tilde{\Psi}_B$. The following lemma shows that $\tilde{\Psi}_B(\cdot)$ converges to $\Phi(\cdot/\sigma_I)$ uniformly in probability.

Lemma 1. *Suppose Assumptions 2–6 hold. Then, $\sup_{\xi\in\mathbb{R}} |\tilde{\Psi}_B(\xi) - \Phi(\xi/\sigma_I)| \xrightarrow{p} 0$.*

Lemma 1 is a key step to show the following theorem, which states that $\tilde{\tau}_\alpha$ for any $0 < \alpha < 1$ well approximates $\sigma_I z_\alpha/\sqrt{n}$.

Theorem 2 (Asymptotic Minimum Half-Width and Exact Coverage of Sample Shrinkage Basic Bootstrap (SSB)). *Suppose Assumptions 2–6 hold. Then, for any $0 < \alpha < 1$, $\sqrt{n}\tilde{\tau}_\alpha = \sigma_I z_\alpha + o_p(1)$. Moreover, the SSB CI has an asymptotically exact coverage at the $1-\alpha$ level, that is, $P\{\psi(F^c) \in [\overline{Y}_{R_0}(\hat{F}_0) - \tilde{\tau}_{1-\alpha/2}, \overline{Y}_{R_0}(\hat{F}_0) - \tilde{\tau}_{\alpha/2}]\} \to 1-\alpha$ as $n \to \infty$, where P is taken with respect to the joint randomness from the data, bootstrapping and simulation runs.*

Analyzing the quantile shrinkage CI is more straightforward. Theorem 3 follows directly from that $\sqrt{n}\hat{\tau}_\beta \xrightarrow{p} (\sigma_I^2 + \sigma^2/p)^{1/2} z_\beta$ and Proposition 3; thus, its proof is omitted.

Theorem 3 (Asymptotic Minimum Half-Width and Exact Coverage of Quantile Shrinkage Basic Bootstrap (QSB)). *Suppose Assumptions 2–6 hold. Then, for any $0 < \alpha < 1$, $(1-\hat{c})\hat{\tau}_\alpha = \sigma_I z_\alpha + o_p(1)$. Moreover, the QSB CI has an asymptotically exact coverage at the $1-\alpha$ level, that is, $P\{\psi(F^c) \in [\overline{Y}_{R_0}(\hat{F}_0) - (1-\hat{c})\hat{\tau}_{1-\alpha/2}, \overline{Y}_{R_0}(\hat{F}_0) - (1-\hat{c})\hat{\tau}_{\alpha/2}]\} \to 1-\alpha$ as $n \to \infty$, where P is taken with respect to the joint randomness from the data, bootstrapping, and simulation runs.*

In contrast, without shrinkage, the BB CI exhibits overcoverage.

Theorem 4 (Asymptotic Limit and Overcoverage of BB). *Suppose Assumptions 2–6 hold. Then, $\sqrt{n}(\overline{Y}_R(\hat{F}_b) - \overline{Y}_{R_0}(\hat{F}_0)) \mid \hat{F}_0 \Rightarrow N(0,\sigma^2/p + \sigma_I^2)$ in probability. Consequently, the BB CI gives asymptotic coverage of*

$$P\{\psi(F^c) \in [\overline{Y}_{R_0}(\hat{F}_0) - \hat{\tau}_{1-\alpha/2}, \overline{Y}_{R_0}(\hat{F}_0) - \hat{\tau}_{\alpha/2}]\} \to \Phi_{\sigma_I^2}(\Phi^{-1}_{\sigma^2/p+\sigma_I^2}(1-\alpha/2)) - \Phi_{\sigma_I^2}(\Phi^{-1}_{\sigma^2/p+\sigma_I^2}(\alpha/2))$$

as $n \to \infty$, where P is taken with respect to the joint randomness from the data, bootstrapping and simulation runs, and $\Phi_a(\cdot)$ denotes the distribution function of $N(0, a)$.

Theorem 4 stipulates that without shrinkage, $\overline{Y}_R(\hat{F}_b)$ centered at $\overline{Y}_{R_0}(\hat{F}_0)$ satisfies a CLT that has a larger variance $\sigma^2/p + \sigma_I^2$ than that in the CLT of the shrinkage estimator σ_I^2. This inflated variance arises from the simulation noise in computing the resample estimator that contributes σ^2/p to the overall variance. As a result, if we use the quantiles of this naive resample estimator to construct a BB CI, we get an asymptotic coverage probability of $\Phi_{\sigma_I^2}(\Phi^{-1}_{\sigma^2/p+\sigma_I^2}(1-\alpha/2)) - \Phi_{\sigma_I^2}(\Phi^{-1}_{\sigma^2/p+\sigma_I^2}(\alpha/2))$, which is the probability content of $N(0,\sigma_I^2)$ between the $\alpha/2$ and $(1-\alpha/2)$th quantiles of a mismatched $N(0,\sigma_p^2 + \sigma_I^2)$ variable and is strictly greater than $1-\alpha$. This reaffirms our motivation to use shrinkage to resolve the coverage issue brought by IU when applying quantile-based bootstraps.

We close this section with an asymptotic guarantee in directly using the bootstrap variance estimates and the CLT to construct CIs in the form of (11).

Theorem 5 (Validity of Variance Bootstrap and Direct Use of Central Limit Theorem). *Under Assumptions* 1, 3, 4, *and* 6, *define* $\hat{V}$ *and* $\hat{W}$ *as in Theorem* 3. *Then,*

$$\left[\overline{Y}_{R_0}(\hat{F}_0) - z_{1-\alpha/2}\sqrt{\max\{\hat{V} - \hat{W}/R\}}, \overline{Y}_{R_0}(\hat{F}_0) + z_{1-\alpha/2}\sqrt{\max\{\hat{V} - \hat{W}/R\}}\right]$$

is an asymptotically exact $(1-\alpha)$*-level CI for* $\psi(F^c)$.

We refer to the CI in Theorem 5 as the *N* method in the remainder of the paper.

5. Implementation Issues and Robustifying the Shrinkage Methods

Recall that the shrinkage factor, c, cannot be computed exactly and is replaced with its sample estimate $\hat{c}$ in Algorithms 1 and 2. However, when IU is relatively small compared with the simulation error variance, then c tends to be close to one. In this case, the probability that the estimated $\hat{c}$ equals one may be significant. If we plug $\hat{c} = 1$ in the place of c in (12), then $\hat{Y}_R(\hat{F}_b) = \overline{\overline{Y}}_R$ for all $b = 1, 2, \ldots, B$, which makes the resulting shrinkage CIs have width of zero.

To avoid this, we propose to replace c with a lower confidence bound of a $1-\beta$ CI for c instead of its point estimate $\hat{c}$. This makes the shrinkage to be less aggressive and more robust to the estimation error in $\hat{c}$. The resulting shrinkage CIs are wider than when $\hat{c}$ is used. Algorithm 3 details how the lower confidence bound, c_{lo}, is computed from the same simulation outputs used to compute $\hat{c}$ with no additional simulation effort. Essentially, Algorithm 3 bootstraps the simulation runs made at $\hat{F}_1, \hat{F}_2, \ldots, \hat{F}_B$ to compute $\hat{c}_h$ for $h = 1, 2, \ldots, H$, then returns the $\beta/2$ empirical quantile of $\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_H$ as c_{lo}.

Algorithm 3 (Computing the Lower Confidence Bound for c)

Input: $\{Y_1(\hat{F}_b), Y_2(\hat{F}_b), \ldots, Y_R(\hat{F}_b)\}_{1 \le b \le B}$ from Algorithm 1
1: **for** $h = 1, 2, \ldots, H$ **do**
2: **for** $l = 1, 2, \ldots, B$ **do**
3: Sample b'_l from $\{1, 2, \ldots, B\}$ with equal probabilities.
4: Compute $\hat{c}_h$ from $\{Y_1(\hat{F}_{b'_l}), Y_2(\hat{F}_{b'_l}), \ldots, Y_R(\hat{F}_{b'_l})\}_{1 \le l \le B}$ from (14).
5: Let c_{lo} be the $\beta/2$ empirical quantile of $\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_H$.

In all our experiments in the following section, we adopt c_{lo} instead of $\hat{c}$ to compute SSB, SSP, QSB, and QSP CIs.

6. Computational Experiments

We present three experiment results to compare the performances of the proposed shrinkage approaches with those of direct bootstrap and the CLT-based CI in (11). That is, we compare BB, PB, N, SSB, SSP, QSB, and QSP.

Section 6.1 presents an M/M/1/k queueing simulation model, where the focus is on comparing the correctness of coverage guarantees and the tightness of CIs. Section 6.2 examines an M/G/1/k queueing example to illustrate the advantage of nonparametric approaches when the parametric family assumption is incorrect. In Section 6.3, we compare the shrinkage approaches with metamodel-based approaches when the number of input distributions is large using a Jackson network example. In all sections, we use $H = 1{,}000$ and $\beta = 0.05$ to compute c_{lo}. All source codes for the experiments in this section are made available at https://github.com/INFORMSJoC/2022.0044-1 (Song et al. 2023).

6.1. Steady-State *M*/*M*/1/*k*

We simulated an $M/M/1/10$ system, where the number of jobs in the system is initialized by sampling from its steady-state distribution. The time in system (TIS) of the first-entering customer after initialization is calculated via Lindley's equation and returned as a simulation output. Two traffic intensities, $\rho = 0.7$ and 0.9, are tested. For each ρ, four sample sizes n = 100, 400, 1,000, and 4,000 are examined. Recall that Theorems 2–5 prescribe $R_0 = \omega(n)$ and $R/n \to p$ for some $p > 0$. To match this, we set $R_0 = \lceil n^{1.1} \rceil$ and test $p = 0.05, 0.2,$ and 0.5. We adopt B = 1,000 in all experiments.

Table 1 shows the average coverage probabilities and widths for nominal 95% CIs for mean TIS computed from 1,000 macroruns. For each combination of $\rho, n,$ and R, the method that shows the closest coverage probability to 95% is marked bold. The average CI widths for BB and PB are consistently larger than *N* in all cases, whereas all shrinkage methods' widths match those of *N* closely. BB and PB show overcoverage when $p = 0.05$ because the simulation error is large relative to the input error. However, as p increases, PB still overcovers, whereas BB show significant undercoverage for n = 100 and n = 400, which worsens under heavier traffic ($\rho = 0.9$). BB's coverage matches 0.95 most closely for $\rho = 0.9$ when $(n, R) = (400, 80)$ and $(n, R) = (1000, 200)$.

Table 1. Average Coverage Probabilities (c) and Widths (w) of 95% CIs for $M/M/1/10$ System from 1,000 Macroruns

ρ	n	R	c_{lo}	BB	PB	N	SSB	SSP	QSB	QSP	
0.7	100	5	0.50	1.000	1.000	0.911	0.883	0.973	0.921	**0.972**	c
				5.35	5.35	2.21	2.71	2.71	2.71	2.71	w
		20	0.30	0.953	0.991	0.927	0.838	0.956	0.872	**0.951**	c
				3.35	3.35	2.26	2.36	2.36	2.36	2.36	w
		50	0.17	0.893	0.971	0.908	0.822	**0.943**	0.834	**0.943**	c
				2.75	2.75	2.28	2.30	2.30	2.30	2.30	w
	400	20	0.52	1.000	1.000	0.901	0.905	0.974	0.930	**0.969**	c
				2.72	2.72	1.09	1.32	1.32	1.32	1.32	w
		80	0.31	0.976	0.993	**0.943**	0.886	0.963	0.906	0.965	c
				1.67	1.67	1.11	1.16	1.16	1.16	1.16	w
		200	0.18	0.924	0.982	0.922	0.847	**0.947**	0.863	0.942	c
				1.36	1.36	1.11	1.13	1.13	1.13	1.13	w
	1,000	50	0.52	1.000	1.000	0.923	**0.941**	0.978	0.962	0.979	c
				1.73	1.73	0.69	0.83	0.83	0.83	0.83	w
		200	0.31	0.984	0.998	**0.950**	0.905	0.966	0.932	0.968	c
				1.05	1.05	0.69	0.72	0.72	0.72	0.72	w
		500	0.17	0.947	0.986	0.931	0.891	0.947	0.902	**0.950**	c
				0.86	0.86	0.69	0.71	0.71	0.71	0.71	w
	4,000	200	0.53	1.000	1.000	0.920	0.934	0.979	**0.958**	0.974	c
				0.86	0.86	0.34	0.41	0.41	0.41	0.41	w
		800	0.31	0.985	0.999	0.942	0.909	**0.956**	0.922	0.957	c
				0.52	0.52	0.34	0.36	0.36	0.36	0.36	w
		2,000	0.18	**0.951**	0.980	0.913	0.883	0.945	0.889	0.939	c
				0.43	0.43	0.34	0.35	0.35	0.35	0.35	w
0.9	100	5	0.47	0.99	1.000	0.859	0.823	0.96	0.848	**0.951**	c
				6.65	6.65	3.05	3.53	3.53	3.53	3.53	w
		20	0.26	0.922	0.986	0.878	0.804	**0.933**	0.819	0.928	c
				4.29	4.29	3.11	3.19	3.19	3.19	3.19	w
		50	0.14	0.837	**0.969**	0.873	0.774	0.927	0.787	0.928	c
				3.60	3.60	3.11	3.11	3.11	3.11	3.11	w
	400	20	0.45	0.997	1.000	0.909	0.913	0.966	0.925	**0.963**	c
				3.58	3.58	1.75	1.97	1.97	1.97	1.97	w
		80	0.23	**0.952**	0.989	0.901	0.864	0.945	0.875	0.942	c
				2.35	2.35	1.76	1.81	1.81	1.81	1.81	w
		200	0.12	0.910	0.967	0.925	0.878	**0.948**	0.884	0.947	c
				2.01	2.01	1.76	1.77	1.77	1.77	1.77	w
	1,000	50	0.45	0.997	0.999	0.92	0.907	0.958	0.918	**0.956**	c
				2.31	2.31	1.15	1.28	1.28	1.28	1.28	w
		200	0.22	**0.957**	0.986	0.917	0.883	**0.943**	0.896	0.942	c
				1.53	1.53	1.15	1.19	1.19	1.19	1.19	w
		500	0.11	0.928	0.962	0.924	0.896	**0.940**	0.902	0.935	c
				1.31	1.31	1.15	1.16	1.16	1.16	1.16	w
	4,000	200	0.44	0.999	0.999	0.919	0.927	0.955	0.944	0.954	c
				1.17	1.17	0.59	0.65	0.65	0.65	0.65	w
		800	0.22	0.974	0.994	0.933	0.916	0.964	0.922	**0.961**	c
				0.77	0.77	0.59	0.61	0.61	0.61	0.61	w
		2,000	0.11	0.955	0.969	0.935	0.919	0.949	0.925	**0.950**	c
				0.67	0.67	0.59	0.59	0.59	0.59	0.59	w

Notes. Standard errors of all CI widths are less than 0.02. For each instance, the method exhibiting the closest coverage probability to 0.95 is marked bold.

However, in both cases, as R increases the coverage worsens instead of improving. This implies that the idealized BB CI without simulation error would in fact undercover, but we observe reasonable coverages for smaller R values because of the simulation error convolution. The normal CI (N) exhibits consistent undercoverage across almost all conditions and perform particularly poorly for $\rho = 0.9$ and $n = 100$. Both BB and N's undercoverage behaviors can be explained by asymmetry of the output function of this example, which we further investigate at the end of this section.

Among the shrinkage bootstrap methods, SSB shows significant undercoverage in all cases. QSB mostly undercovers except when n is 1,000 or 4,000 and $p = 0.05$ for $\rho = 0.7$. This is not surprising, considering the poor performance of BB in this example. Because SSB and QSB are "shrunk" versions of BB, they are expected show worse

coverage when BB undercovers. Conversely, SSP and QSP show most robust performance across all experiment settings. For n = 1,000 and n = 4,000, all methods show improved performance than when n is smaller, as the bootstrap distribution more closely approximates a normal distribution. Nevertheless, SSP and QSP still exhibit advantages over PB and N matching the 95% coverage target more closely with tighter CIs.

Table 1 also reports the average c_{lo} of all 1,000 macroruns for each parameter setting. For each n, as R increases, the simulation error shrinks and c_{lo} decreases as a consequence. Typically, SSP and QSP are most effective when c_{lo} is larger as the overcoverage and the CI width of PB can be significantly reduced. When c_{lo} is smaller, there is not much room to reduce overcoverage by shrinkage and SSP and QSP may even undercover; see when $(n, R) = (100, 50)$. Recall from Section 5 that c_{lo} can be computed within each macrorun. Thus, a user may decide whether to adopt SSP/QSP over PB by evaluating c_{lo}.

To provide further insights on the performance difference among BB and PB and their shrunk variants, we present a histogram of average TIS calculated at 1,000 bootstrap samples when $\rho = 0.9, n = 100,$ and R = 50 in Figure 1. The symbols, △ and •, respectively represent the mean TIS at F^c $(= \psi(F^c))$ and $\overline{Y}_{R_0}(\hat{F}_0)$, whereas the solid and dashed lines represent BB and PB CIs, respectively. Because the bootstrap sample means are right-skewed for this example, the upper bound of BB is closer to $\overline{Y}_{R_0}(\hat{F}_0)$ than that of PB. As a result, when $\overline{Y}_{R_0}(\hat{F}_0) < \psi(F^c)$ and their difference is sufficiently large, BB may fail to cover $\psi(F^c)$ as shown in Figure 1. Such a phenomenon is more likely to happen for higher traffic intensity (more skewed) and smaller n. Conversely, the PB CI benefits from the existence of a monotonic transformation and covers $\psi(F^c)$. The observed sknewness in Figure 1 also explains why N, which produces a symmetric CI, undercovers.

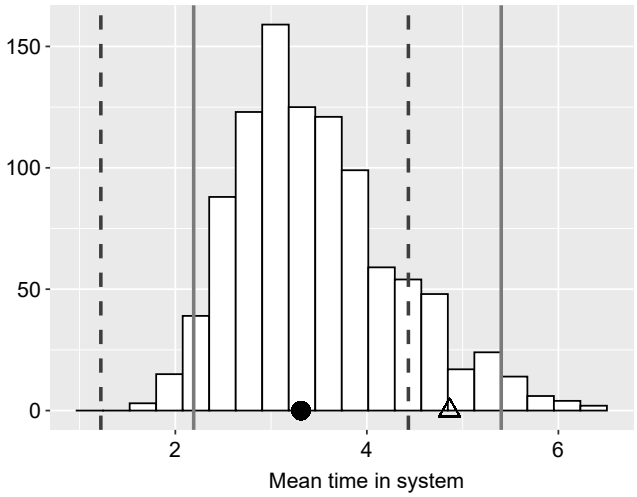
6.2. Finite-Horizon $M/G/1/k$

In this section, we present the simulation results from an $M/G/1/10$ system starting empty. The CIs are constructed for the mean TIS for the first 20 jobs. Service times were generated from the bimodal distribution adopted in Ghosh and Lam (2019): $0.3\text{Beta}(2, 6) + 0.7\text{Beta}(6, 2)$. We set the arrival rate to be $\lambda \approx 1.17$, which results in $\rho = 0.7$. The mean TIS of the first 20 jobs estimated from 10^6 replications is 1.17.

Table 2 shows the average coverage probabilities and widths of nominal 95% CIs for mean TIS computed from 1,000 macroruns under two different real-world sample sizes, n = 400 and n = 1,000, whereas B = 1,000 is adopted for all experiments. For each n, we set $R_0 = \lceil n^{1.1} \rceil$ and $R = np$ for p = 0.05 and 0.2. For each combination of n and R, the method that exhibits coverage closest to 95% is marked bold.

In Table 2, pPBe refers to the parametric bootstrap that adopts fitted exponential distributions for interarrival and service time distributions, where the latter is clearly fitted to a wrong distribution family. The poor coverage of pPBe demonstrates that it can be detrimental to take a parametric approach with a wrong distribution family. Observe that SSP and QSP show excellent coverage probabilities in all experiment settings outperforming N. As in the $M/M/1/k$ case, BB, SSB, and QSB show undercoverage due to the skewness of the mean TIS distribution.

Figure 1. (Color online) Histogram of Average TIS of 1,000 Bootstrap Samples



Notes. △ represents the mean TIS at F^c and • is $\overline{Y}_{R_0}(\hat{F}_0)$; the dashed/solid lines show BB/PB CIs. Observe that BB does not cover △.

Table 2. Average Coverage Probabilities (c) and Widths (w) of 95% CIs for $M/G/1/10$ System from 1,000 Macroruns

n	R	c_{lo}	pPBe	BB	PB	N	SSB	SSP	QSB	QSP	
400	20	0.34	0.005	0.990	0.998	0.942	0.917	**0.957**	0.925	**0.957**	c
			0.48	0.51	0.51	0.31	0.34	0.34	0.34	0.34	w
	80	0.14	0.000	0.941	0.975	0.941	0.917	**0.948**	0.918	0.947	c
			0.35	0.37	0.37	0.32	0.32	0.32	0.32	0.32	w
1,000	50	0.34	0.000	0.985	0.997	0.939	0.932	**0.958**	0.939	**0.958**	c
			0.30	0.32	0.32	0.20	0.21	0.21	0.21	0.21	w
	200	0.14	0.000	0.970	0.977	0.949	0.939	**0.950**	0.938	0.952	c
			0.22	0.24	0.24	0.20	0.20	0.20	0.20	0.20	w

Notes. Standard errors of all CI widths are less than 0.02. For each instance, the method exhibiting the closest coverage probability to 0.95 is marked bold.

Moreover, SSP and QSP exhibit more dramatic reductions in the overcoverage and the width compared with PB when c_{lo} is larger.

6.3. Capacitated Jackson Network

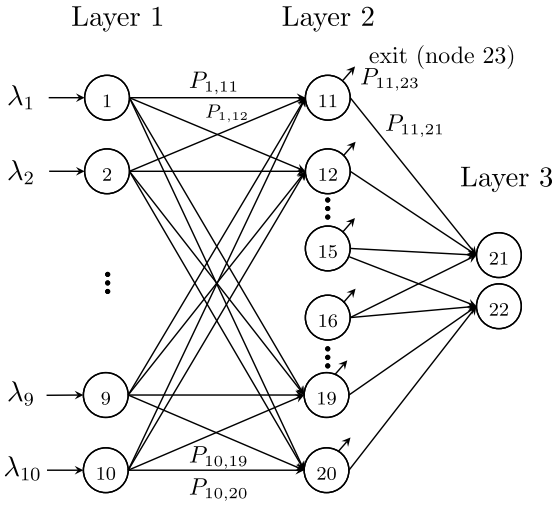
In this section, we consider a capacitated Jackson network, where each node has a single server with capacity of 100 and exponential service time distributions with known mean = 1. The network structure has a layer of m input nodes followed by a layer of m intermediate nodes and an output layer of two nodes. Two network forms are considered: $m = 4$ and $m = 10$. Jobs arrive at the m nodes in Layer 1, then are routed randomly to Layer 2 and (possibly) subsequently to Layer 3. The structure of the network for $m = 10$ is shown in Figure 2.

We assume that the arrival process at each node in Layer 1 is Poisson with average rate across the m nodes of $\rho = 0.7$. Once a job is finished at a node in Layer 1 or 2, the job is sent to the next node according to the routing probabilities in Tables OS.1 and OS.2 in Section OS.2 of the online supplement. Some jobs exit the network after Layer 2 instead of being routed to one of the two processing nodes in Layer 3. For all nonparametric methods, we assume that the interarrival time distributions and routing probabilities are unknown and estimated by empirical distributions from observations.

Each replication is initialized with an empty system. After a warm-up of 100 jobs, output is computed as the average cumulative waiting time of the next 10 jobs whose services were completed without balking for each of R simulation replications.

If one is willing to make parametric assumptions on the input distributions, the metamodel-based approach by Barton et al. (2014) and Xie et al. (2014) has been shown to work well for a problem with a small to moderate number of input distribution parameters. We apply the metamodel-based approach to the Jackson network example assuming the correct input distribution families are known—a particularly favorable condition. For this example, the parameter vector, $\boldsymbol{\xi}$, of the joint input distribution consisting of arrival rates and routing probabilities is either 22 ($m = 4$) or 112 ($m = 10$) dimensional.

Figure 2. Three-Layer Jackson Network Structure



As in Barton et al. (2014), we fit a stochastic Gaussian process (GP) with the squared exponential correlation function after running replications at a set of design points in the space of $\boldsymbol{\xi}$. This is implemented in R using the `mlegp` package. For large numbers of parameters, the design construction algorithm in Barton et al. (2014) is not practical. An alternate space-filling design used here is discussed in Section OS.3 of the online supplement. Let the maximum likelihood estimators (MLEs) computed from the input data and the bth bootstrap sample drawn from empirical distributions be $\boldsymbol{\xi}_0$ and $\boldsymbol{\xi}_b$, respectively. Moreover, we denote the posterior GP mean by $m(\boldsymbol{\xi})$ and the bth order statistic of $m(\boldsymbol{\xi}_1), m(\boldsymbol{\xi}_2), \ldots, m(\boldsymbol{\xi}_B)$ by $m_{(b)}$. The two metamodel-bootstrap methods we compare are as follows:

- Basic metamodel CI (BM): $[2\overline{Y}_{R_0}(\boldsymbol{\xi}_0) - m_{(\lceil B(1-\alpha/2)\rceil)}, 2m(\boldsymbol{\xi}_0) - m_{(\lceil B(\alpha/2)\rceil)}]$
- Percentile metamodel CI (PM): $[m_{(\lceil B(\alpha/2)\rceil)}, m_{(\lceil B(1-\alpha/2)\rceil)}]$

Note that $\overline{Y}_{R_0}(\boldsymbol{\xi}_0)$ is the sample average of R_0 replications run using $\boldsymbol{\xi}_0$ as the input parameter vector, which is analogous to $\overline{Y}_{R_0}(\hat{F}_0)$ in the nonparametric approaches. A version of PM was proposed to provide a Bayesian credible interval (Xie et al. 2014); in their work, the PM interval was constructed using stochastic samples from the posterior GP instead of the posterior means to incorporate the prediction error. In our experiments, we define BM and PM without prediction error, which eliminates the explicit inclusion of simulation error variance on the CIs and produces narrower CIs.

In Table 3, we compare the empirical coverage probabilities and widths of 95% CIs constructed by each method from 1,000 macroruns for the $m = 4$ capacitated Jackson network. To compare the computational costs of bootstrap versus metamodel approaches, we report the CPU time in minutes (Xeon E5-2680 processors) per macrorun: $T_{\text{CPU,D}}$ is the average time for computing all nonparametric bootstrap and N CIs and $T_{\text{CPU,MM}}$ is the average time for computing BM and PM CIs.

Two sets of sample sizes are tested: (i) 100 for interarrival time distributions and 1,000 for routing counts and (ii) 400 and 4,000, respectively. To evaluate the coverage, the true mean waiting time is computed from 4×10^6 Monte Carlo simulations using the true input distributions. For BM and PM, the GP model is fitted from R replications made at each of K space-filling design points, where $K = 220$ for $m = 4$ and $K = 1{,}000$ for $m = 10$. The choices of K roughly follow the guidance for a design size that is ten times the number of parameters for GP models Jones et al. (1998). Because the metamodel evaluations do not require simulations, BM and PM evaluate the fitted metamodel at $B = 1{,}000$ bootstrap samples to construct the CIs, for both $m = 4$ and $m = 10$. Conversely, to compute the CIs for BB, PB, N, SSB, SSP, QSB, and QSP, we adopt $B = 220$ for $m = 4$ and $B = 1{,}000$ for $m = 10$ to make the computational cost of these methods and metamodel-based methods comparable.

Table 3 shows several interesting features. First, as in Sections 6.1 and 6.2, BB and PB overcover when there is substantial simulation error, namely, when c_{lo} is higher. Except for when the real-world sample size is 100/1,000 and $R = 40$, SSP and QSP provide good coverages with 58%–75% interval widths of BB/PB/N. We continue to observe that BB's performance is inferior to PB. Even for metamodel-based intervals, PM outperforms BM. This is likely due to the asymmetric distribution of the output statistic and is reduced with larger sample size.

As R increases, c_{lo} decreases, and then the overcoverage by BB, PB, and N is reduced, and undercoverage for SSP and QSP begins to appear. However, when the real-world sample sizes increase from 100/1,000 to 400/4,000, IU is reduced making the simulation error relatively large for the same R. Consequently, c_{lo} increases

Table 3. Average Coverage Probabilities (c) and Widths (w) of 95% CIs for the $m = 4$ Jackson Network from 1,000 Macroruns

n	R	c_{lo}	$T_{\text{CPU,D}}$	$T_{\text{CPU,MM}}$	BB	PB	N	SSB	SSP	QSB	QSP	BM	PM	
100/1,000	10	0.34	0.3	1.5	0.984	0.996	0.998	0.902	0.935	0.922	**0.944**	0.917	0.941	c
					3.14	3.14	3.21	2.08	2.08	2.08	2.08	2.22	2.22	w
	40	0.15	1.4	1.9	0.929	**0.963**	**0.963**	0.871	0.920	0.880	0.922	0.901	0.924	c
					2.22	2.22	2.27	1.88	1.88	1.88	1.88	2.14	2.14	w
	160	0.05	4.3	4.6	0.878	0.942	0.933	0.860	0.934	0.862	0.933	0.890	**0.953**	c
					1.92	1.92	1.97	1.83	1.83	1.83	1.83	2.20	2.21	w
400/4,000	10	0.51	0.3	1.1	1.000	1.000	1.000	**0.956**	0.977	0.967	0.977	0.937	0.960	c
					2.72	2.72	2.77	1.35	1.35	1.35	1.35	1.16	1.16	w
	40	0.34	0.9	2.0	0.992	1.000	0.999	0.922	0.963	0.940	0.964	**0.942**	0.962	c
					1.57	1.57	1.61	1.04	1.04	1.04	1.04	1.09	1.09	w
	160	0.15	3.2	4.1	0.936	0.974	0.973	0.897	0.947	0.902	**0.949**	0.939	0.979	c
					1.12	1.12	1.14	0.95	0.95	0.95	0.95	1.14	1.14	w

Notes. Two sets of sample sizes are tested: (i) 100 for interarrival time distributions and 1,000 for routing counts and (ii) 400 and 4,000, respectively. The CPU times, $T_{\text{CPU,D}}$ and $T_{\text{CPU,MM}}$, are reported in minutes.

Table 4. Average Coverage Probabilities (c) and Widths (w) of 95% CIs for the $m = 10$ Jackson Network from 1,000 Macroruns

w	c_{lo}	$T_{\text{CPU,D}}$	$T_{\text{CPU,MM}}$	BB	PB	N	SSB	SSP	QSB	QSP	BM	PM	
5	0.61	1.3	240	0.999	1.000	1.000	0.790	0.956	0.909	0.958	0.786	**0.948**	c
				1.98	1.98	1.99	0.77	0.77	0.77	0.77	0.69	0.69	w
50	0.44	1.5	290	0.998	0.998	0.998	0.898	0.923	0.926	**0.941**	0.919	0.879	c
				1.34	1.34	1.35	0.75	0.75	0.75	0.75	0.78	0.78	w
400	0.20	3.8	260	0.962	0.964	0.974	0.925	0.912	0.926	0.922	**0.946**	0.877	c
				1.36	1.36	1.37	1.09	1.09	1.09	1.09	1.21	1.21	w

Notes. We set $R = 10$ and the real-world sample sizes of 100 for the interarrival times and 1,000 for routing counts are adopted. Here, w represents the number of completed customers we observe after the warm-up within each replication. The CPU times, $T_{\text{CPU,D}}$ and $T_{\text{CPU,MM}}$, are reported in minutes.

and the shrinkage-based methods show improved performances. The performances of BM and PM are also improved for the larger real-world sample case. This is because the sampling distribution of the MLEs are more concentrated as the sample size increases, which reduces the response surface complexity and improves the goodness-of-fit of the metamodel.

Last, BM and PM show good coverage, which can be attributed to that we do not include stochastic error term when constructing BM/PM CI, and thus it is much less sensitive to overcoverage from simulation error. However, metamodel-based methods have two disadvantages. First, they require parametric distributions for input probability models. Although a flexible family based on Bayesian mixtures was proposed by Xie et al. (2021), parametric assumptions are not always appropriate, and misspecification can lead to significant errors—as seen in the $M/G/1/k$ examples. Second, without special model simplifications, the computation time needed to fit GP metamodels is $O(k^3)$ per optimization iteration. That is, because multiple iterations are typically required to find the MLEs for the GP model, the computational cost can be significant. These two issues are more apparent in the results for the larger network presented in Table 4.

Table 4 compares the empirical coverage probabilities and widths of 95% CIs from 1,000 macroruns for the $m = 10$ capacitated Jackson network. This network has 10 arrival rates and 102 routing probabilities for a total of 112 parameters for the GP metamodel. The larger network has greater simulation error. Additionally, we control the length of each replication in this set of experiments by observing w number of completed customers waiting times after the warmup. The larger w is, the smaller the simulation error is. Again, for each experiment, the method that has the coverage closest to 95% is marked bold.

Table 4 shows less satisfactory performances for metamodel-based CIs. PM performs well when $c_{lo} = 0.61$, producing the tightest CI and the closest coverage to 95%. However, the coverage deteriorates with smaller simulation error, indicating perhaps difficulty from mean reversion of the GP prediction, which increases when the fitted GP model has a small simulation error variance estimate. BM shows improved coverage as c_{lo} decreases. This phenomenon is not predictable, however, and does not support BM as an effective method in general. Conversely, the increase in execution time for the metamodel-based approach is substantial, 70–200 times the execution time for the nonparametric approaches, with more than 80% of the total associated with metamodel fitting.

As for the $m = 4$ case, for large simulation error ($c_{lo} = 0.61$) the SSP and QSP methods produce good coverage with CIs that are from 38% to 39% of the width of the BB, PB, or N CIs. For moderate simulation error ($c_{lo} = 0.44$), SSP and QSP coverage is still close to 95%, with CI width approximately 56% of BB, PB, and N CIs, which continue to overcover. However, coverage for the shrinkage methods deteriorates for the low simulation error case ($c_{lo} = 0.20$), and CI widths are at approximately 80% of the BB, PB, and N CIs.

Overall, the Jackson network results show good performance for SSP, QSP, and PM when there is substantial simulation error. However, for problems with many parameters, PM is computationally inefficient. We conclude that the shrinkage methods can perform as well as or better than metamodel-based methods even in the case the correct parametric families are assumed known. The advantage can be expected to grow as the number of input models increases, because getting an adequate fit becomes more difficult and the computational effort associated with fitting the GP increases.

7. Conclusion

Direct bootstrap resampling provides a nonparametric characterization of IU free of assumptions on the parametric families, but it has significant overcoverage arising from the simulation stochastic error when the simulation effort is limited. To remedy the overcoverage, we designed two new bootstrap CI approaches that

characterize and compensate for simulation error in a computationally efficient way without expending overwhelming simulation effort to make the simulation error negligible. Both approaches use shrinkage strategies to properly scale down the statistical variability implied by the CI due to the simulation error. The first approach, sample shrinkage, shrinks bootstrap sample means toward a suitable average, and the second approach, quantile shrinkage, directly adjusts the empirical quantiles of the bootstrap sample means in forming the CI.

Unlike the classical analyses on shrinkage, we investigate the case when the simulation errors are heteroscedastic and provide guidance on the choices of bootstrap sample size B and the number of replications at each bootstrap, R. We show that our shrinkage strategies give rise to CIs with widths that are on par with the asymptotically tight normal CI implied by the central limit theorem even when the shrinkage factor is estimated via simulations. Moreover, to robustify the shrinkage CIs against the estimation error of the shrinkage factor, we propose to construct a lower confidence bound for the shrinkage factor via bootstrapping the simulation outputs without additional simulation runs. The resulting lower bound can also be used as a guide to decide when to apply shrinkage over the classical bootstrap. From our empirical studies, a rule of thumb may be that when the lower bound is greater than 0.2, the user can benefit from applying the shrinkage percentile bootstrap CIs. We also conducted several computational experiments, which demonstrated (i) the reduced CI widths and overcoverage of the shrinkage methods, (ii) the advantage of nonparamteric approaches when input distributions are improperly characterized, and (iii) the advantage over the parametric metamodel-based approach when the number of parameters is large and the data size is limited.

Methods to address the accuracy of the bootstrap when the resulting simulation output is near a system boundary, analyses to refine accuracy at even higher-order levels, and other approaches to further reduce simulation effort are all areas for future research.

Acknowledgments

The authors thank the anonymous referees for thoughtful comments and suggestions, Madhura Mundale for help with coding the Jackson network example, and Wei Xie for sharing the codes to generate the experiment design for metamodeling used in Barton et al. (2014). Computations for this research were performed on the Pennsylvania State University's Institute for Computational and Data Sciences' Roar supercomputer, with assistance from Carrie Brown, Research Innovation with Scientists and Engineers Team, and Graham Lockard, Smeal Research Instruction & Information Technology Group.

References

Barton RR, Schruben LW (1993) Uniform and bootstrap resampling of input distributions. Evans GW, Mollaghasemi M, Russell EC, Biles WE, eds. *Proc. Winter Simulation Conf.* (IEEE, Piscataway, NJ), 503–508.

Barton RR, Schruben LW (2001) Resampling methods for input modeling. Peters BA, Smith JS, Medeiros DJ, Rohrer MW, eds. *Proc. Winter Simulation Conf.* (IEEE, Piscataway, NJ), 372–378.

Barton RR, Lam H, Song E (2018) Revisiting direct bootstrap resampling for input model uncertainty. Rabe M, Juan AA, Mustafee N, Skoogh A, Jain S, Johansson B, eds. *Proc. Winter Simulation Conf.* (IEEE, Piscataway, NJ), 1635–1645.

Barton RR, Nelson BL, Xie W (2014) Quantifying input uncertainty via simulation confidence intervals. *INFORMS J. Comput.* 26(1):74–87.

Cheng RCH, Holland W (1997) Sensitivity of computer simulation experiments to errors in input data. *J. Statist. Comput. Simulations* 57(1–4):219–241.

Cheng RCH, Holland W (2004) Calculation of confidence intervals for simulation output. *ACM Trans. Modeling Comput. Simululations* 14:344–362.

Davison AC, Hinkley DV (1997) *Bootstrap Methods and Their Application.* Cambridge Series in Statistical and Probabilistic Mathematics (Cambridge University Press, Cambridge, UK).

Efron B (1987) *The Jackknife, the Bootstrap, and Other Resampling Plans* (Society for Industrial Mathematics, Philadelphia).

Efron B, Tibshirani RJ (1994) *An Introduction to the Bootstrap* (CRC Press, Boca Raton, FL).

Flynn TN, Peters TJ (2004) Use of the bootstrap in analysing cost data from cluster randomised trials: Some simulation results. *BMC Health Services Res.* 4:33.

Ghosh S, Lam H (2019) Robust analysis in stochastic simulation: Computation and performance guarantees. *Oper. Res.* 67(1):232–249.

Glynn PW, Lam H (2018) Constructing simulation output intervals under input uncertainty via data sectioning. Rabe M, Juan AA, Mustafee N, Skoogh A, Jain S, Johansson B, eds. *Proc. Winter Simulation Conf.* (IEEE, Piscataway, NJ), 1551–1562.

Hampel FR (1974) The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* 69(346):383–393.

Jones D, Schonlau M, Welch W (1998) Efficient global optimization of expensive black-box functions. *J. Global Optim.* 13:455–492.

Kosorok MR (2007) *Introduction to Empirical Processes and Semiparametric Inference* (Springer Science & Business Media, Boston).

Lam H, Qian H (2016) The empirical likelihood approach to simulation input uncertainty. Roeder TMK, Frazier PI, Szechtman R, Zhou E, Huschka T, Chick SE, eds. *Proc. Winter Simulation Conf.* (IEEE, Piscataway, NJ), 791–802.

Lam H, Qian H (2018) Subsampling variance for input uncertainty quantification. Rabe M, Juan AA, Mustafee N, Skoogh A, Jain S, Johansson B, eds. *Proc. Winter Simulation Conf.* (IEEE, Piscataway, NJ), 1611–1622.

Lin Y, Song E, Nelson B (2015) Single-experiment input uncertainty. *J. Simulations* 9(3):249–259.

McIntyre J, Stefanski LA (2011) Density estimation with replicate heteroscedastic measurements. *Ann. Institute Statist. Math.* 63(1):81–99.
Morgan LE, Nelson BL, Titman AC, Worthington DJ (2019) Detecting bias due to input modelling in computer simulation. *Eur. J. Oper. Res.* 279(3):869–881.
Ng ESW, Grieve R, Carpenter JR (2013) Two-stage nonparametric bootstrap sampling with shrinkage correction for clustered data. *Stata J.* 13(1):141–164.
Song E (2021) Sequential bayesian risk set inference for robust discrete optimization via simulation. Preprint, January 19, https://arxiv.org/abs/2101.07466.
Song E, Nelson BL (2015) Quickly assessing contributions to input uncertainty. *IIE Trans.* 47(9):893–909.
Song E, Lam H, Barton RR (2023) A shrinkage approach to improve direct bootstrap resampling under input uncertainty. http://dx.doi.org/10.1287/ijoc.2022.0044.cd, https://github.com/INFORMSJoC/2022.0044.
Song E, Nelson BL, Pegden CD (2014) Advanced tutorial: Input uncertainty quantification. Tolk A, Diallo S, Ryzhov I, Yilmaz L, Buckley S, Miller J, eds. *Proc. Winter Simulation Conf.* (IEEE, Piscataway, NJ), 162–176.
Van der Vaart AW (2000) *Asymptotic Statistics*, vol. 3 (Cambridge University Press, Cambridge, UK).
Wang H, Ng SH, Zhang X (2020) A Gaussian process based algorithm for stochastic simulation optimization with input distribution uncertainty. Bae K-H, Feng B, Kim S, Lazarova-Molnar S, Zheng Z, Roeder T, Thiesing R, eds. *Proc. Winter Simulation Conf.* (IEEE, Piscataway, NJ), 2899–2910.
Xie W, Nelson BL, Barton RR (2014) A Bayesian framework for quantifying uncertainty in stochastic simulation. *Oper. Res.* 62(6):1439–1452.
Xie W, Nelson BL, Barton RR (2016) Multivariate input uncertainty in output analysis for stochastic simulation. *ACM Trans. Modeling Comput. Simulation* 27(1):1–22.
Xie W, Li C, Wu Y, Zhang P (2021) A nonparametric Bayesian framework for uncertainty quantification in stochastic simulation. *SIAM/ASA J. Uncertainty Quantification* 9(4):1527–1552.
Yi Y, Xie W (2017) An efficient budget allocation approach for quantifying the impact of input uncertainty in stochastic simulation. *ACM Trans. Modeling Comput. Simulations* 27(4):25.