Journal of Data Science 22 (1), 25–44
January 2024

DOI: 10.6339/23-JDS1092
Statistical Data Science

Impact of Bias Correction of the Least Squares Estimation on Bootstrap Confidence Intervals for Bifurcating Autoregressive Models

Tamer Elbayoumi[1] and Sayed Mostafa[1,*]

[1]*Department of Mathematics & Statistics, North Carolina A&T State University, USA*

Abstract

The least squares (LS) estimator of the autoregressive coefficient in the bifurcating autoregressive (BAR) model was recently shown to suffer from substantial bias, especially for small to moderate samples. This study investigates the impact of the bias in the LS estimator on the behavior of various types of bootstrap confidence intervals for the autoregressive coefficient and introduces methods for constructing bias-corrected bootstrap confidence intervals. We first describe several bootstrap confidence interval procedures for the autoregressive coefficient of the BAR model and present their bias-corrected versions. The behavior of uncorrected and corrected confidence interval procedures is studied empirically through extensive Monte Carlo simulations and two real cell lineage data applications. The empirical results show that the bias in the LS estimator can have a significant negative impact on the behavior of bootstrap confidence intervals and that bias correction can significantly improve the performance of bootstrap confidence intervals in terms of coverage, width, and symmetry.

Keywords *BAR model; binary trees; cell-lineage; maternal correlation*

1 Introduction

The bifurcating autoregressive (BAR) model is commonly used to model binary tree-structured data, depicted in Figure 1, that appear in many applications, most famously cell-lineage applications (e.g., Cowan, 1984; Hawkins et al., 2009; Kimmel and Axelrod, 2005; Sandler et al., 2015). The first-order BAR [BAR(1)] model was first introduced and studied by Cowan and Staudte (1986) for modeling cell-lineage data. This model can be seen as an extension of the first-order autoregressive [AR(1)] model where each line of descent is modeled as an AR(1) process with the observations from the two sibling cells who share the same parent being correlated. In practice, the BAR(1) model is used to explain the progression of single-cell proliferation (c.f., Kimmel and Axelrod, 2005).

Several studies have considered the problem of estimation of the BAR model parameters (see, e.g., Cowan and Staudte, 1986; Huggins, 1995; Bui and Huggins, 1999; Huggins and Basawa, 1999, 2000; Zhou and Basawa, 2005; Terpstra and Elbayoumi, 2012; Elbayoumi and Terpstra, 2016, among many others). Cowan and Staudte (1986) introduced and studied maximum likelihood (ML) estimators for the model coefficients and the correlation between errors in the BAR(1) model assuming the normality of the model errors. Huggins and Basawa (1999) studied the asymptotic properties of the ML estimators for the BAR(p) model. On the other hand, Zhou

[*]Corresponding author. Email: sabdelmegeed@ncat.edu.

Received October 7, 2022; Accepted February 19, 2023

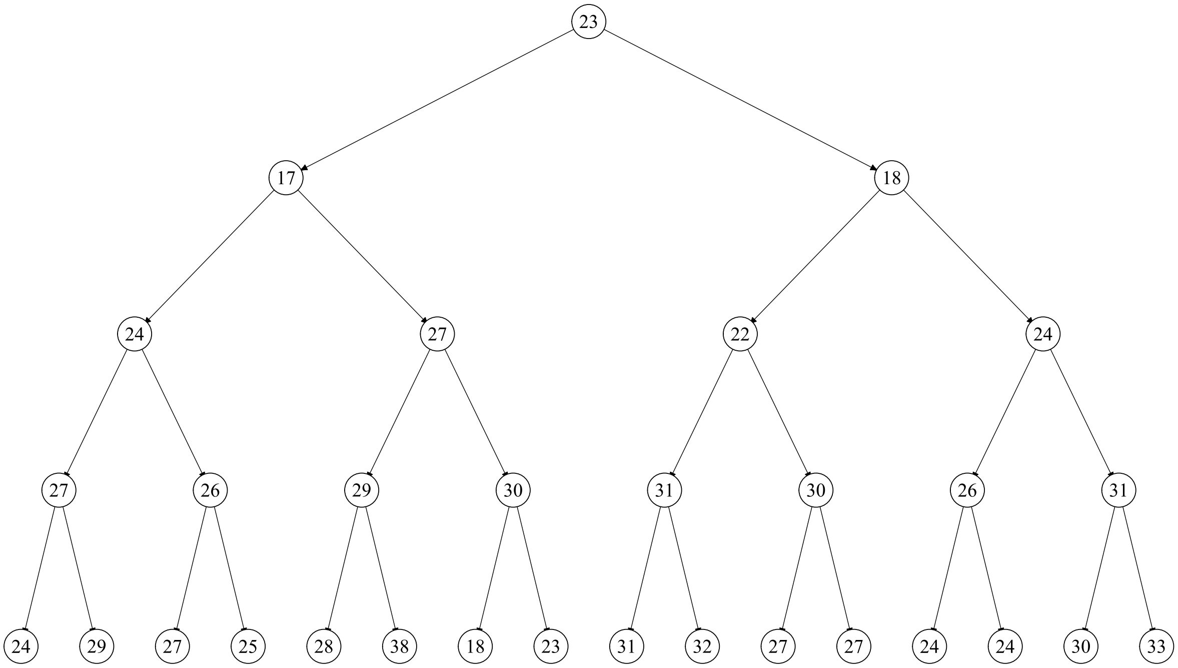

Figure 1: Lifetimes (in minutes) of E. Coli cells. The tree is made using data from Cowan and Staudte (1986).

and Basawa (2005) focused on the least squares (LS) estimation for BAR(p) models and did not make distributional assumptions about the model errors. In their study, Zhou and Basawa (2005) defined LS estimators for the model coefficients, a consistent estimator for the correlation between model errors, and derived the limiting distribution of the LS estimators. Elbayoumi and Mostafa (2021b) investigated the finite sample properties of the LS estimators of the BAR(1) model coefficients. They showed that the finite sample bias of these estimators can be quite large, especially for parameter values close to the boundary points, which might make inferences based on these estimators to be inaccurate. Elbayoumi and Mostafa (2021b) introduced two methods for correcting the bias in the LS estimators of the BAR model autoregressive parameter, namely, bootstrap bias correction and bias correction through linear bias functions. The results of Elbayoumi and Mostafa (2021b) showed that both the linear-bias-correcting estimator and the bootstrap bias-correcting estimators can be quite effective in reducing the bias of the LS estimator and that the bootstrap estimators are more effective near the boundaries of the range of the autoregressive parameter.

In this paper, we focus on the construction of bootstrap confidence intervals for the BAR model. Specifically, we assess the impact of bias in LS estimators on the behavior of bootstrap confidence intervals for the autoregressive parameter (ϕ_1, the target parameter) in the BAR(1) model. We conduct an extensive empirical study to evaluate the impact of different types of bias correction of the LS estimator for the BAR(1) model on the performance of several confidence interval procedures.

The rest of the paper is organized as follows. In Section 2, we give an overview of the BAR(1) model and the LS estimation of model parameters, demonstrate the bias in the LS estimation of the autoregressive parameter ϕ_1, and outline two bootstrap bias correction approaches. Section 3 introduces several types of bootstrap confidence intervals (with and without bias correction) for ϕ_1. In Section 4, we report and discuss the results of an extensive empirical study based on both Monte Carlo simulations and two real data applications. The paper is concluded with some discussions in Section 5.

2 The BAR Model and LS Estimation Bias

Let $X_1, X_2, \ldots, X_n$ denote the random variables corresponding to the observations on a perfect binary tree with g generations. The initial observation X_1 corresponds to generation 0, while the observations $X_{2^i}, X_{2^i+1}, \ldots, X_{2^{i+1}-1}$ correspond to the 2^i observations in generation $i, i = 1, 2, \ldots, g$. Note that the sample size $n = 2^{g+1} - 1$. Mathematically, the BAR(1) model is given by

$$X_t = \mathbf{Z}_t'\boldsymbol{\phi} + \varepsilon_t, \text{ for all } t \geqslant 2, \tag{1}$$

where X_t is an observed value of some quantitative characteristic at time t, $\mathbf{Z}_t' = (1, X_{[t/2]})$, where $X_{[t/2]}$ denotes the mother of X_t for all $t \geqslant 2$ and $[u]$ denotes the largest integer less than or equal to u. The parameter vector $\boldsymbol{\phi}' = (\phi_0, \phi_1)$ is a vector of unknown model coefficients, where ϕ_0 is the intercept and ϕ_1 denotes the autoregressive parameter (aka the inherited effect or the maternal correlation). It is assumed that $\phi_1 \in (-1, 1)$, which implies that the autoregressive process is stationary. The errors $(\varepsilon_{2t}, \varepsilon_{2t+1})$ are independently and identically distributed (iid) according to some joint distribution F. Here, $(\varepsilon_{2t}, \varepsilon_{2t+1})$ have zero mean vector and variance-covariance matrix given by

$$\boldsymbol{\Sigma}_{\varepsilon_t} = \begin{pmatrix} 1 & \theta \\ \theta & 1 \end{pmatrix} \sigma^2, \tag{2}$$

where θ denotes the linear correlation between ε_{2t} and ε_{2t+1} (θ is also known as the environmental effect or the sister-sister correlation given their common mother) and σ^2 is the variance of the errors. Furthermore, it is assumed that the pairs $(\varepsilon_{2t}, \varepsilon_{2t+1})$ and $(\varepsilon_{2s}, \varepsilon_{2s+1})$ are independent for $s \neq t$. The rationale for this correlation structure is that sister cells grow in the same environment, especially early in their lives. Therefore, two distinct kinds of correlation are expected: (i) the environmental correlation between siblings; and (ii) the maternal correlation due to the effects inherited from the mother. On the other hand, other distant relatives, such as cousins, share less in their environment and, thus, it is reasonable to assume that their environmental effects are independent.

2.1 LS Estimation of the BAR Model

Zhou and Basawa (2005) derived the following LS estimators of the BAR(1) model parameters:

$$\hat{\phi}_1^{\text{LS}} = \left[\sum_{t=1}^{m} U_t(X_t - \bar{X})\right]\left[\sum_{t=1}^{m}(X_t - \bar{X})^2\right]^{-1},$$

$$\hat{\phi}_0^{\text{LS}} = \bar{U} - \hat{\phi}_1^{LS}\bar{X},$$

where $m = (n-1)/2$, $U_t = (X_{2t} + X_{2t+1})/2$, $\bar{X} = m^{-1}\sum_{t=1}^{m} X_t$, and $\bar{U} = m^{-1}\sum_{t=1}^{m} U_t$. The consistent estimators of σ^2 and θ are given by $\hat{\sigma}^2 = (n-3)^{-1}\sum_{t=2}^{n} \hat{\varepsilon}_t^2$ and

$$\hat{\theta}^{\text{LS}} = \hat{\sigma}^{-2} m^{-1} \sum_{t=1}^{m} \hat{\varepsilon}_{2t}\hat{\varepsilon}_{2t+1}, \tag{3}$$

where $\hat{\varepsilon}_t = X_t - \mathbf{Z}_t'\hat{\boldsymbol{\phi}}$. Zhou and Basawa (2005) also showed that the joint limiting distribution of the LS estimators of the BAR(1) model coefficients is given by:

$$\sqrt{n}(\hat{\boldsymbol{\phi}}^{\text{LS}} - \boldsymbol{\phi}) \xrightarrow{d} N(0, \sigma^2(1+\theta)\boldsymbol{A}^{-1}), \tag{4}$$

where

$$\boldsymbol{A} = \begin{pmatrix} 1 & \frac{\phi_0}{(1-\phi_1)} \\ \frac{\phi_0}{(1-\phi_1)} & \frac{\sigma^2}{(1-\phi_1^2)} + (\frac{\phi_0}{1-\phi_1})^2 \end{pmatrix}.$$

There are several other parameters that can be studied under the BAR(1) model but are of less interest. Examples may include (i) the unconditional sibling-sibling correlation: $\rho = \phi_1^2 + (1-\phi_1^2)\theta$, (ii) the cousin-cousin correlation: $\rho\phi_1^2$, and (iii) the grandmother-granddaughter correlation: ϕ_1^2. Since $\hat{\theta}$ is computed from the residuals, estimates for the three parameters (i–iii) are easily obtainable once ϕ_1 is estimated. Therefore, we confine our attention to the estimation of the main parameter ϕ_1.

2.2 The Bias in LS Estimation of the BAR Model

Elbayoumi and Mostafa (2021b) studied the bias in the LS estimation of the BAR(1) autoregressive coefficient, ϕ_1. They derived the asymptotic bias of the LS estimator and demonstrated the finite sample bias using Monte Carlo simulations. Their results showed that the bias of $\hat{\phi}_1^{\text{LS}}$ can be quite significant, especially for small sample sizes. Specifically, they showed that for small sample sizes ($n = 31$) and when ϕ_1 is near zero (-0.15 to $+0.15$), the bias of $\hat{\phi}_1^{\text{LS}}$ can be as large as 5% to 60% (as θ moves from -1 to $+1$) of the true value of ϕ_1. For other values of ϕ_1 larger than -0.3, the relative bias of $\hat{\phi}_1^{\text{LS}}$ continues to be considerably large as long as θ is away from -1. The bias remains substantial in most cases, sometimes reaching 30%, for moderate sample sizes, e.g., $n = 63$. We refer the reader to Fig. 3 in Elbayoumi and Mostafa (2021b) for a full depiction of the bias in the LS estimator of the BAR(1) model autoregressive coefficient.

2.3 Bootstrap Bias Correction Methods for the BAR Model

The bootstrap technique has been widely used for bias correction in many statistical estimation problems. For instance, the model-based bootstrap has been applied to correct the bias in the LS estimation of the AR(1) model coefficient (e.g., Berkowitz and Kilian, 2000; Liu-Evans and Phillips, 2012). Recently, this method was extended by Elbayoumi and Mostafa (2021b) to correct the bias in the LS estimation of the BAR(1) model coefficient. For completeness and later reference, we briefly describe three model-based bootstrap bias correction methods for the BAR(1) model.

We start by rewriting the BAR(1) model given in Eq. (1) as follows: for $t \geqslant 1$,

$$X_{2t} = \phi_0 + \phi_1 X_t + \varepsilon_{2t},$$

$$X_{2t+1} = \phi_0 + \phi_1 X_t + \varepsilon_{2t+1},$$

where X_{2t} and X_{2t+1} are the observations for the two siblings branching from X_t. Given the observed binary tree $\mathbf{X}^n = (X_1, X_2, \ldots, X_n)$, compute the LS estimates of the BAR(1) model coefficients; $\hat{\phi}_0^{\text{LS}}$ and $\hat{\phi}_1^{\text{LS}}$, as described in Section 2.1, and obtain the residuals $\hat{\varepsilon}_{2t} = X_{2t} - (\hat{\phi}_0^{\text{LS}} + \hat{\phi}_1^{\text{LS}} X_t)$ and $\hat{\varepsilon}_{2t+1} = X_{2t+1} - (\hat{\phi}_0^{\text{LS}} + \hat{\phi}_1^{\text{LS}} X_t)$ for all $t \geqslant 1$ and the centered residuals $\tilde{\varepsilon}_t = \hat{\varepsilon}_t - (n-1)^{-1}\sum_{i=2}^{n}\hat{\varepsilon}_i$ for all $t \geqslant 2$.

Algorithm 1 Single Bootstrap Bias-Corrected LS Estimation for ϕ_1 (SBC).

```
Input
Observed tree: $\mathbf{X}^n = (X_1, X_2, \ldots, X_n)$
LS estimates of the BAR(1) model coefficients: $\hat{\phi}_0^{\mathrm{LS}}$ and $\hat{\phi}_1^{\mathrm{LS}}$
Centered residuals: $\tilde{\varepsilon}_t = \hat{\varepsilon}_t - (n-1)^{-1} \sum_{i=2}^{n} \hat{\varepsilon}_i,\ t \geqslant 2$
Number of bootstrap resamples: $B$
for each $b \leftarrow 1$ to $B$ do
    Set $X_{1,b}^* = X^0$, the last observation in an initial binary tree[§], of size $n^0 = 31$[§§], whose first
    observation is $X_1$ in the original observed tree
    for each $j \leftarrow 1$ to $m = (n-1)/2$ do
        Sample with replacement one pair $(\tilde{\varepsilon}_{2j,b}^*, \tilde{\varepsilon}_{2j+1,b}^*)$ from among the pairs
        $\{(\tilde{\varepsilon}_{2t}, \tilde{\varepsilon}_{2t+1});\ t \geqslant 1\}$
        Compute
            $X_{2j,b}^* = \hat{\phi}_0^{\mathrm{LS}} + \hat{\phi}_1^{\mathrm{LS}} X_{j,b}^* + \tilde{\varepsilon}_{2j,b}^*,$
            $X_{2j+1,b}^* = \hat{\phi}_0^{\mathrm{LS}} + \hat{\phi}_1^{\mathrm{LS}} X_{j,b}^* + \tilde{\varepsilon}_{2j+1,b}^*$
    end for
    Build the bootstrap tree $\mathbf{X}_b^{*n} = (X_{1,b}^*, X_{2,b}^*, \ldots, X_{n,b}^*)$
    Compute the LS estimate $\hat{\phi}_{1_b}^*$ from $\mathbf{X}_b^{*n}$
end for
Obtain the bootstrap estimate of the bias of $\hat{\phi}_1^{\mathrm{LS}}$ as $\hat{\beta}_{\hat{\phi}_1^{\mathrm{LS}}} = \frac{1}{B} \sum_{b=1}^{B} (\hat{\phi}_{1_b}^* - \hat{\phi}_1^{\mathrm{LS}})$
Output: The single bootstrap bias-corrected LS estimate of $\phi_1$:
    $\hat{\phi}_1^{\mathrm{SBC}} = \hat{\phi}_1^{\mathrm{LS}} - \hat{\beta}_{\hat{\phi}_1^{\mathrm{LS}}}$    (5)
```

[§] In each bootstrap replicate, the initial tree is generated from a BAR(1) model whose coefficients are the observed LS estimates $\hat{\phi}_0^{\mathrm{LS}}$ and $\hat{\phi}_1^{\mathrm{LS}}$ and whose errors are randomly sampled in pairs from among the centered residuals. We note that the alternative approach of using X_1 from the observed tree as the first observation in all bootstrap trees may add artificial correlation among the bootstrap trees. This issue can become more pronounced in higher-order BAR(p), $p > 1$, models where $2^p - 1$ initial observations would be needed to start the bootstrap process. In such case, our approach would simply take the $2^p - 1$ observations from the end of the initial tree.

[§§] Any suitable initial tree size, n^0, can be used. The suggested $n^0 = 31$ seems to produce good balance between computing time and stability of results.

The single bootstrap bias correction procedure for the autoregressive coefficient in the BAR(1) model is presented in Algorithm 1. We note that the amount of bias in the corrected estimate $\hat{\phi}_1^{\mathrm{SBC}}$ is of order $O(n^{-2})$, whilst the original estimator $\hat{\phi}_1^{\mathrm{LS}}$ has a bias of order $O(n^{-1})$. The double bootstrapping approach involves iterating the single bootstrap procedure twice. This technique was shown to improve the effectiveness of the bootstrap bias correction in the context of the AR(1) model (e.g., Hall, 1992; Lee and Young, 1999; Shi, 1992; Chang and Hall, 2015) and in the context of the BAR(1) model (See, Elbayoumi and Mostafa, 2021b). In Algorithm 2, we describe the double bootstrap approach for correcting the bias in the LS estimator of ϕ_1 under the BAR(1) model. One obvious drawback in this approach appears in its computational cost. This has led to the proposal of a modified double bootstrap bias correction algorithm that is more computationally efficient than the double bootstrap bias correction algorithm. The

Algorithm 2 Double Bootstrap Bias-Corrected LS Estimation for ϕ_1 (DBC).

```
Input:
Observed tree: $\mathbf{X}^n = (X_1, X_2, \ldots, X_n)$
LS estimates of the BAR(1) model coefficients: $\hat{\phi}_0^{\mathrm{LS}}$ and $\hat{\phi}_1^{\mathrm{LS}}$
Centered residuals: $\tilde{\varepsilon}_t = \hat{\varepsilon}_t - (n-1)^{-1} \sum_{i=2}^{n} \hat{\varepsilon}_i$, $t \geqslant 2$
Number of phase 1 bootstrap resamples: $B_1$
Number of phase 2 bootstrap resamples: $B_2$
1:  for each $b \leftarrow 1$ to $B_1$ do
2:      Set $X_{1,b}^* = X^0$, the last observation in an initial binary tree, of size $n^0 = 31$, whose first
        observation is $X_1$, the first observation in $\mathbf{X}^n$
3:      for each $j \leftarrow 1$ to $m = (n-1)/2$ do
4:          Sample with replacement one pair $(\tilde{\varepsilon}_{2j,b}^*, \tilde{\varepsilon}_{2j+1,b}^*)$ from among the pairs
            $\{(\tilde{\varepsilon}_{2t}, \tilde{\varepsilon}_{2t+1});\ t \geqslant 1\}$.
5:          Compute
                $X_{2j,b}^* = \hat{\phi}_0^{\mathrm{LS}} + \hat{\phi}_1^{\mathrm{LS}} X_{j,b}^* + \tilde{\varepsilon}_{2j,b}^*,$
                $X_{2j+1,b}^* = \hat{\phi}_0^{\mathrm{LS}} + \hat{\phi}_1^{\mathrm{LS}} X_{j,b}^* + \tilde{\varepsilon}_{2j+1,b}^*$
6:      end for
7:      Build the bootstrap tree $\mathbf{X}_b^{*n} = (X_{1,b}^*, X_{2,b}^*, \ldots, X_{n,b}^*)$
8:      Compute the first-phase bootstrap LS estimates $\hat{\phi}_{0_b}^*$ and $\hat{\phi}_{1_b}^*$ from $\mathbf{X}_b^{*n}$
9:      Obtain the first-phase bootstrap residuals
            $\hat{\varepsilon}_{2t,b}^* = X_{2t,b}^* - (\hat{\phi}_{0_b}^* + \hat{\phi}_{1_b}^* X_{t,b}^*)$ and $\hat{\varepsilon}_{2t+1,b}^* = X_{2t+1,b}^* - (\hat{\phi}_{0_b}^* + \hat{\phi}_{1_b}^* X_{t,b}^*);\ t \geqslant 1$
10:     Obtain the centered bootstrap residuals $\tilde{\tilde{\varepsilon}}_{t,b}^* = \hat{\varepsilon}_{t,b}^* - (n-1)^{-1} \sum_{i=2}^{n} \hat{\varepsilon}_{i,b}^*;\ t \geqslant 2$
11:     for each $k \leftarrow 1$ to $B_2$ do
12:         Repeat steps 2-6 on $\mathbf{X}_b^{*n}$ with $\hat{\phi}_{0_b}^*$, $\hat{\phi}_{1_b}^*$ and the centered residuals $\tilde{\tilde{\varepsilon}}_{t,b}^*$
13:         Build the second-phase bootstrap tree $\mathbf{X}_{k,b}^{**n} = (X_{1,kb}^{**}, X_{2,kb}^{**}, \ldots, X_{n,kb}^{**})$
14:         Compute the second-phase bootstrap LS estimate $\hat{\phi}_{1_{kb}}^{**}$ from $\mathbf{X}_{k,b}^{**n}$
15:     end for
16: end for
17: Obtain the single bootstrap estimate of the bias of $\hat{\phi}_1^{\mathrm{LS}}$ as:
        $\hat{\beta}_{\hat{\phi}_1^{\mathrm{LS}}} = \frac{1}{B_1} \sum_{b=1}^{B_1} (\hat{\phi}_{1_b}^* - \hat{\phi}_1^{\mathrm{LS}})$
18: Obtain the double bootstrap bias adjustment factor as:
        $\hat{\gamma}_{\hat{\phi}_1^{\mathrm{LS}}} = \hat{\beta}_{\hat{\phi}_1^{\mathrm{LS}}} - \frac{1}{B_1 B_2} \sum_{b=1}^{B_1} \sum_{k=1}^{B_2} (\hat{\phi}_{1_{kb}}^{**} - \hat{\phi}_{1_b}^*)$
Output: The double bootstrap bias-corrected LS estimate for $\phi_1$: $\hat{\phi}_1^{\mathrm{DBC}} = \hat{\phi}_1^{\mathrm{LS}} - \hat{\beta}_{\hat{\phi}_1^{\mathrm{LS}}} - \hat{\gamma}_{\hat{\phi}_1^{\mathrm{LS}}}$
```

modification is known as the fast-double bootstrap and it was studied by Ouysse (2013). The fast-double bootstrap algorithm uses one bootstrap resample in phase 2 (i.e., $B_2 = 1$) from each bootstrap sample created in phase 1. This is unlike the double bootstrap algorithm which draws B_2 bootstrap resamples within each phase bootstrap sample. While the double bootstrap algorithm has a computational cost of order $O(B_1 B_2)$, the fast-double bootstrap algorithm's cost is of order $O(B_1)$. Despite this significant difference in computational cost, Elbayoumi and Mostafa (2021b) noted that the bias correction performance of the two algorithms is quite similar.

3 Confidence Intervals for the BAR Model

Now, we present various types of confidence interval procedures for the BAR(1) autoregressive coefficient ϕ_1. Specifically, we describe several bootstrap confidence intervals for ϕ_1 based on the LS estimator $\hat{\phi}_1$ with and without bias correction. We start with a Wald-type confidence interval for ϕ_1 based on the asymptotic distribution of $\hat{\phi}_1$ and a plug-in estimate of its asymptotic standard error. We call it the asymptotic confidence interval and use it as a benchmark when evaluating the performance of the bootstrap confidence intervals.

3.1 Asymptotic Confidence Interval

Recall from (4) that $\hat{\phi}_1$ has asymptotic normal distribution with mean ϕ_1 and variance given by

$$\mathrm{Var}(\hat{\phi}_1^{\mathrm{LS}}) = \frac{1}{n}(1+\theta)(1-\phi_1^2).$$

Using this asymptotic distribution, a $100(1-\alpha)\%$ asymptotic CI for ϕ_1 is given by

$$\hat{\phi}_1^{\mathrm{LS}} \pm z_{\alpha/2}\widehat{se}_{\mathrm{asy}}(\hat{\phi}_1^{\mathrm{LS}}),$$

where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ is the $(1-\alpha/2)$ percentile under the standard normal distribution and

$$\widehat{se}_{\mathrm{asy}}(\hat{\phi}_1^{\mathrm{LS}}) = \sqrt{\frac{1}{n}(1+\hat{\theta}^{\mathrm{LS}})(1-\hat{\phi}_1^{\mathrm{LS}^2})},$$

with $\hat{\theta}^{\mathrm{LS}}$ as given by (3).

It should be noted that this confidence interval can be highly impacted by the bias in the LS estimator $\hat{\phi}^{\mathrm{LS}}$ through the point estimate and the standard error. Therefore, alternative confidence interval procedures are in demand to mitigate the bias effect.

3.2 Standard Normal Bootstrap Confidence Intervals

Instead of approximating the standard error of $\hat{\phi}^{\mathrm{LS}}$ using a plug-in estimate of the asymptotic standard error, one can use the bootstrap approach to directly estimate the standard error and construct confidence intervals for ϕ. In the following, we present such confidence intervals with and without correction for the bias in $\hat{\phi}^{\mathrm{LS}}$.

- **Uncorrected standard normal bootstrap CI.** Given an observed tree of size n observations, apply the bootstrap approach to obtain B replicates of the estimator $\hat{\phi}^{\mathrm{LS}}$, say; $\hat{\phi}_{1_{(1)}}^{\mathrm{LS}^*}, \hat{\phi}_{1_{(2)}}^{\mathrm{LS}^*}, \ldots, \hat{\phi}_{1_{(B)}}^{\mathrm{LS}^*}$. The bootstrap estimate of the standard error of $\hat{\phi}^{\mathrm{LS}}$ is then obtained as

$$\widehat{se}_{\mathrm{boot}}(\hat{\phi}_1^{\mathrm{LS}}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\phi}_{1_{(b)}}^{\mathrm{LS}^*} - \bar{\hat{\phi}}_1^{\mathrm{LS}^*}\right)^2},$$

with $\bar{\hat{\phi}}_1^{\mathrm{LS}*} = \frac{1}{B}\sum_{b=1}^{B} \hat{\phi}_{1_{(b)}}^{\mathrm{LS}*}$. A $100(1-\alpha)\%$ standard normal bootstrap CI for ϕ_1 is given by

$$\hat{\phi}_1^{\mathrm{LS}} \pm z_{\alpha/2}\widehat{se}_{\mathrm{boot}}(\hat{\phi}_1^{\mathrm{LS}}).$$

Similar to the asymptotic confidence interval, the performance of the above bootstrap confidence interval procedure can deteriorate if $\hat{\phi}^{\mathrm{LS}}$ is substantially biased. The following two confidence interval procedures try to address this issue for this kind of confidence interval procedure.

- **Standard normal bootstrap CI with single bootstrap bias correction.** To reduce the impact of the bias in $\hat{\phi}^{\mathrm{LS}}$ on the performance of the standard normal bootstrap CI, we rely on a bias-corrected version of $\hat{\phi}^{\mathrm{LS}}$ when constructing the CI. More precisely, we replace $\hat{\phi}^{\mathrm{LS}}$ with the single bootstrap bias-corrected LS estimator $\hat{\phi}_1^{\mathrm{SBC}}$ presented in (5). This leads to the following bias-corrected CI procedure:

$$\hat{\phi}_1^{\mathrm{SBC}} \pm z_{\alpha/2}\widehat{se}_{\mathrm{boot}}(\hat{\phi}_1^{\mathrm{SBC}}),$$

where

$$\widehat{se}_{\mathrm{boot}}(\hat{\phi}_1^{\mathrm{SBC}}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\phi}_{1_{(b)}}^{\mathrm{SBC}*} - \bar{\hat{\phi}}_1^{\mathrm{SBC}*}\right)^2},$$

with $\hat{\phi}_{1_{(1)}}^{\mathrm{SBC}*}, \hat{\phi}_{1_{(2)}}^{\mathrm{SBC}*}, \ldots, \hat{\phi}_{1_{(B)}}^{\mathrm{SBC}*}$ being B bootstrap replicates of $\hat{\phi}_1^{\mathrm{SBC}}$, and $\bar{\hat{\phi}}_1^{\mathrm{SBC}*}$ being the mean of these replicates. It is noteworthy that this confidence interval procedure requires two phases of bootstrapping, where the first phase corrects the bias in the LS estimator $\hat{\phi}^{\mathrm{LS}}$ to obtain $\hat{\phi}_1^{\mathrm{SBC}}$ and the second phase estimates the standard error of $\hat{\phi}_1^{\mathrm{SBC}}$.

- **Standard normal bootstrap CI with fast double bootstrap bias correction.** Another bias-corrected standard normal bootstrap CI for ϕ_1 can be built based on the fast double bootstrap bias-corrected LS estimator $\hat{\phi}_1^{\mathrm{FDBC}}$ obtained as the output of Algorithm 2 after setting $B_2 = 1$. The fast double bootstrap approach is applied first to obtain the bias-corrected estimator $\hat{\phi}_1^{\mathrm{FDBC}}$ followed by a single bootstrap of $\hat{\phi}_1^{\mathrm{FDBC}}$ to estimate its standard error. The resulting CI is given by

$$\hat{\phi}_1^{\mathrm{FDBC}} \pm z_{\alpha}\widehat{se}_{\mathrm{boot}}(\hat{\phi}_1^{\mathrm{FDBC}}),$$

with

$$\widehat{se}_{\mathrm{boot}}(\hat{\phi}_1^{\mathrm{FDBC}}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\phi}_{1_{(b)}}^{\mathrm{FDBC}*} - \bar{\hat{\phi}}_1^{\mathrm{FDBC}*}\right)^2},$$

where $\hat{\phi}_{1_{(1)}}^{\mathrm{FDBC}*}, \hat{\phi}_{1_{(2)}}^{\mathrm{FDBC}*}, \ldots, \hat{\phi}_{1_{(B)}}^{\mathrm{FDBC}*}$ are B bootstrap replicates of $\hat{\phi}_1^{\mathrm{FDBC}}$, and $\bar{\hat{\phi}}_1^{\mathrm{FDBC}*}$ is the mean of these replicates.

3.3 Percentile Bootstrap Confidence Interval

- **Uncorrected percentile bootstrap CI.** Given an observed tree of size n observations, apply the bootstrap approach to obtain B replicates of the estimator $\hat{\phi}^{\mathrm{LS}}$, say; $\hat{\phi}_{1_{(1)}}^{\mathrm{LS}*}, \hat{\phi}_{1_{(2)}}^{\mathrm{LS}*}, \ldots, \hat{\phi}_{1_{(B)}}^{\mathrm{LS}*}$. The $100(1-\alpha)\%$ percentile bootstrap CI for ϕ_1 can then be obtained as

$$\left(\hat{\phi}_{1_{\alpha/2}}^{\mathrm{LS}*}, \hat{\phi}_{1_{1-\alpha/2}}^{\mathrm{LS}*}\right),$$

where $\hat{\phi}^{\mathrm{LS}^*}_{1_{\alpha/2}}$ and $\hat{\phi}^{\mathrm{LS}^*}_{1_{1-\alpha/2}}$ are the $(\alpha/2)$ and $(1-\alpha/2)$ percentiles of the empirical distribution of the B bootstrap replicates of $\hat{\phi}^{\mathrm{LS}}$.

- **Percentile bootstrap CI with single bootstrap bias correction.** The bias in the LS estimator $\hat{\phi}^{\mathrm{LS}}$ can significantly impact the bootstrap distribution and, hence, the behavior of the percentile bootstrap confidence interval. A corrected percentile bootstrap CI for ϕ_1 can be obtained from the bootstrap distribution of the single bootstrap bias-corrected estimator $\hat{\phi}^{\mathrm{SBC}^*}$ as follows:
$$\left(\hat{\phi}^{\mathrm{SBC}^*}_{1_{\alpha/2}}, \hat{\phi}^{\mathrm{SBC}^*}_{1_{1-\alpha/2}}\right),$$
where $\hat{\phi}^{\mathrm{SBC}^*}_{1_{\alpha/2}}$ and $\hat{\phi}^{\mathrm{SBC}^*}_{1_{1-\alpha/2}}$ are the $(\alpha/2)$ and $(1-\alpha/2)$ percentiles of the empirical distribution of the bootstrap replicates $\hat{\phi}^{\mathrm{SBC}^*}_{1_{(1)}}, \hat{\phi}^{\mathrm{SBC}^*}_{1_{(2)}}, \ldots, \hat{\phi}^{\mathrm{SBC}^*}_{1_{(B)}}$.
- **Percentile bootstrap CI with fast double bootstrap bias correction.** Similarly, a $100(1-\alpha)\%$ percentile bootstrap CI for ϕ_1 based on the fast double bootstrap bias-corrected estimator $\hat{\phi}^{\mathrm{FDBC}^*}_1$ is given by
$$\left(\hat{\phi}^{\mathrm{FDBC}^*}_{1_{\alpha/2}}, \hat{\phi}^{\mathrm{FDBC}^*}_{1_{1-\alpha/2}}\right),$$
where $\hat{\phi}^{\mathrm{FDBC}^*}_{1_{\alpha/2}}$ and $\hat{\phi}^{\mathrm{FDBC}^*}_{1_{1-\alpha/2}}$ are the $(\alpha/2)$ and $(1-\alpha/2)$ percentiles of the empirical distribution of the bootstrap replicates $\hat{\phi}^{\mathrm{FDBC}^*}_{1_{(1)}}, \hat{\phi}^{\mathrm{FDBC}^*}_{1_{(2)}}, \ldots, \hat{\phi}^{\mathrm{FDBC}^*}_{1_{(B)}}$.
Note that, unlike the uncorrected bootstrap percentile confidence interval procedure which requires only one phase of bootstrapping, the above bias-corrected percentile bootstrap confidence interval procedures require two phases of bootstrapping.
- **Bias-corrected and accelerated bootstrap CI.** Another modification to the bootstrap percentile confidence interval procedure aims to correct the potential bias and skewness in the bootstrap estimate of the sampling distribution of $\hat{\phi}^{\mathrm{LS}}_1$. This procedure is known as the bias-corrected and accelerated (BCa) bootstrap confidence interval and it has been shown to have improved theoretical properties and enhanced performance in practice (See, Efron, 1987). Define
$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{\alpha/2}}{1-\hat{a}(\hat{z}_0 + z_{\alpha/2})}\right) \quad \text{and} \quad \alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{(1-\alpha/2)}}{1-\hat{a}(\hat{z}_0 + z_{(1-\alpha/2)})}\right),$$
where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution,
$$\hat{z}_0 = \Phi^{-1}\left(\frac{1}{B}\sum_{b=1}^{B} I(\hat{\phi}^{\mathrm{LS}^*}_{1_{(b)}} < \hat{\phi}^{\mathrm{LS}}_1)\right)$$
is the bias-correction factor with $I(\cdot)$ being the indicator function, and
$$\hat{a} = \frac{\sum_{i=1}^{n}(\hat{\phi}^{\mathrm{LS}}_{1(i)} - \bar{\hat{\phi}}^{\mathrm{LS}}_1)^3}{6\left(\sum_{i=1}^{n}(\hat{\phi}^{\mathrm{LS}}_{1(i)} - \bar{\hat{\phi}}^{\mathrm{LS}}_1)^2\right)^{3/2}}$$
is the acceleration factor with $\hat{\phi}^{\mathrm{LS}}_{1(i)}$ being the delete-1-jackknife estimate, i.e., the LS estimate with the i-th observation excluded, and $\bar{\hat{\phi}}^{\mathrm{LS}}_1$ is the mean of these jackknife replicates. A $100(1-\alpha)\%$ BCa CI for ϕ_1 in the BAR(1) model is given by
$$\left(\hat{\phi}^{\mathrm{LS}^*}_{1_{\alpha_1}}, \hat{\phi}^{\mathrm{LS}^*}_{1_{\alpha_2}}\right).$$

Note that the BCa procedure nests the delete-1-jackknife within each of the B bootstrap replicates and, therefore, is more computationally intensive than the uncorrected percentile bootstrap confidence interval procedure.

4 Empirical Results

In this section, we report the results of an empirical study that investigates the performance of eight confidence interval procedures presented in the previous section for the autoregressive coefficient ϕ_1 of the BAR(1) model. The empirical study includes both extensive simulations and two real data applications. All computations were performed using R version 4.1.3 (R Core Team, 2022) on a machine with a processor Intel Xeon E5-2699Av4, 2.4 GHz clock speed, 44 cores, and 512 GB memory. Simulations were set up to utilize 42 of the 44 available cores using the "parallel" package. The "bifurcatingr" package was used to generate the bifurcating trees and compute both uncorrected and bias-corrected LS estimators of the BAR model parameters (Elbayoumi and Mostafa, 2021a).

4.1 Simulations

The simulation experiments are designed to (1) assess the impact of bias in the LS estimator on the performance of various confidence interval procedures for the autoregressive coefficient in the BAR(1) model, and (2) investigate the effectiveness of bias correction for enhancing the performance of these confidence interval procedures. The performance of confidence interval procedures is measured by the empirical coverage rate, width, and symmetry of the confidence interval (i.e., approximately equal error rates on both sides).

In our simulations, we generate perfect binary trees from the BAR(1) model given in Eq. (1). The model's intercept ϕ_0 is set equal to 10 in all trees. We account for wide scenarios of maternal and environmental correlation levels by considering 48 possible combinations of

- $\phi_1 = \pm 0.10, \pm 0.35, \pm 0.60, \pm 0.85$, and
- $\theta = \pm 0.30, \pm 0.6, \pm 0.90$.

The model errors $(\varepsilon_{2t}, \varepsilon_{2t+1})$ are generated as iid observations with zero mean vector and variance-covariance matrix $\mathbf{\Sigma}_{\varepsilon_t}$ given by (2). The following distributions are considered:

- Bivariate normal distribution with 3 signal-to-noise ratios: $\sigma = 0.25, 0.5, 1$.
- Bivariate t-distribution with 10 degrees of freedom and $\sigma = 1$ using the function *rmvst()* in the package "fCopulae" (Wuertz et al., 2022).
- Bivariate skew normal distribution with skewness parameter $a = 3$ and $\sigma = 1$ using the function *rmvsnorm()* in the package "fCopulae".

In each simulated tree, the first observation, X_1 (generation 0), is taken as the last observation in an initial binary tree, of the same size as the target tree, whose first observation is $X_1^{\text{initial}} = \phi_0/(1 - \phi_1)$.

Under each of these settings, $m = 500$ BAR(1) trees, of size $n = 63$ (i.e., $g = 5$ generations), are generated, and the LS estimator of the coefficient ϕ_1 is obtained along with the eight confidence intervals described in the previous section. We also considered two other sample sizes $n = 31, 127$ (i.e., $g = 4, 6$). The eight confidence interval procedures analyzed are: asymptotic Wald-Type CI (ASY); uncorrected standard normal bootstrap CI (NORMB); standard normal bootstrap CI with single bootstrap bias correction (NORMB-SBC); standard normal bootstrap CI with fast double bootstrap bias correction (NORMB-FDBC); uncorrected percentile bootstrap CI (PERC); percentile bootstrap CI with single bootstrap bias correction (PERC-SBC);

percentile bootstrap CI with fast double bootstrap bias correction (PERC-FDBC); and bias-corrected and accelerated bootstrap CI (BCa).

In the case of single bootstrap, the number of bootstrap samples is set to $B = 499$. For the fast double bootstrap, we set $B_1 = 499$ ($B_2 = 1$ by definition). The nominal coverage is set to 95% for all confidence interval procedures.

We use four metrics to compare the performance of the eight confidence interval procedures:

- Coverage: % of time the parameter is within the limits of the CI,
- Average Width,
- Lower Significance Rate (SL.L): % of time the parameter is below the lower limit of the CI, and
- Upper Significance Rate (SL.U): % of time the parameter exceeds the upper limit of the CI.

The simulation results are presented in Figures 2–6 for the 3 error distributions under the

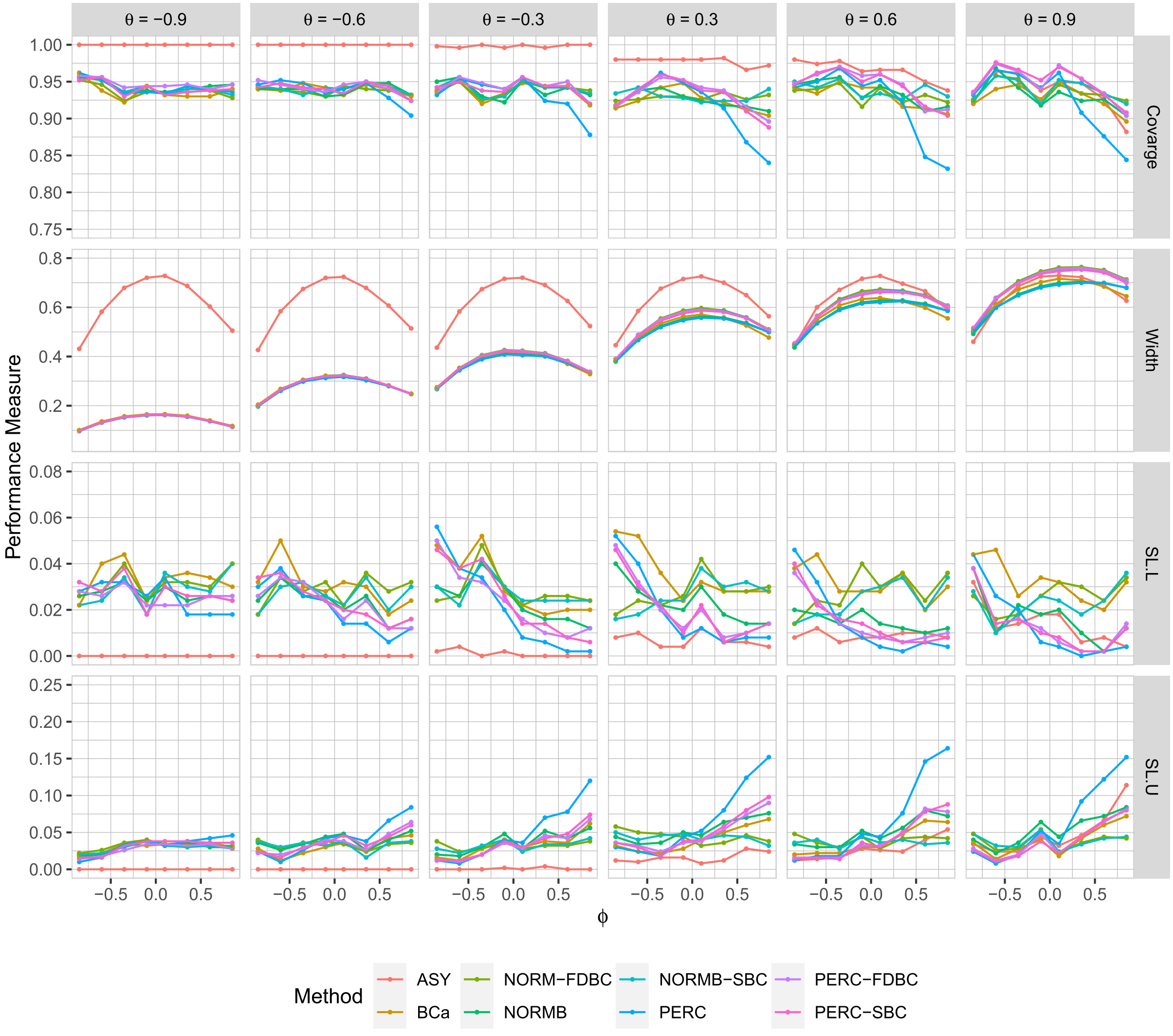

Figure 2: Performance of eight confidence interval procedures of ϕ_1 based on tree size $n = 63$ under the bivariate normal distribution for the errors ($\sigma = 0.25$).

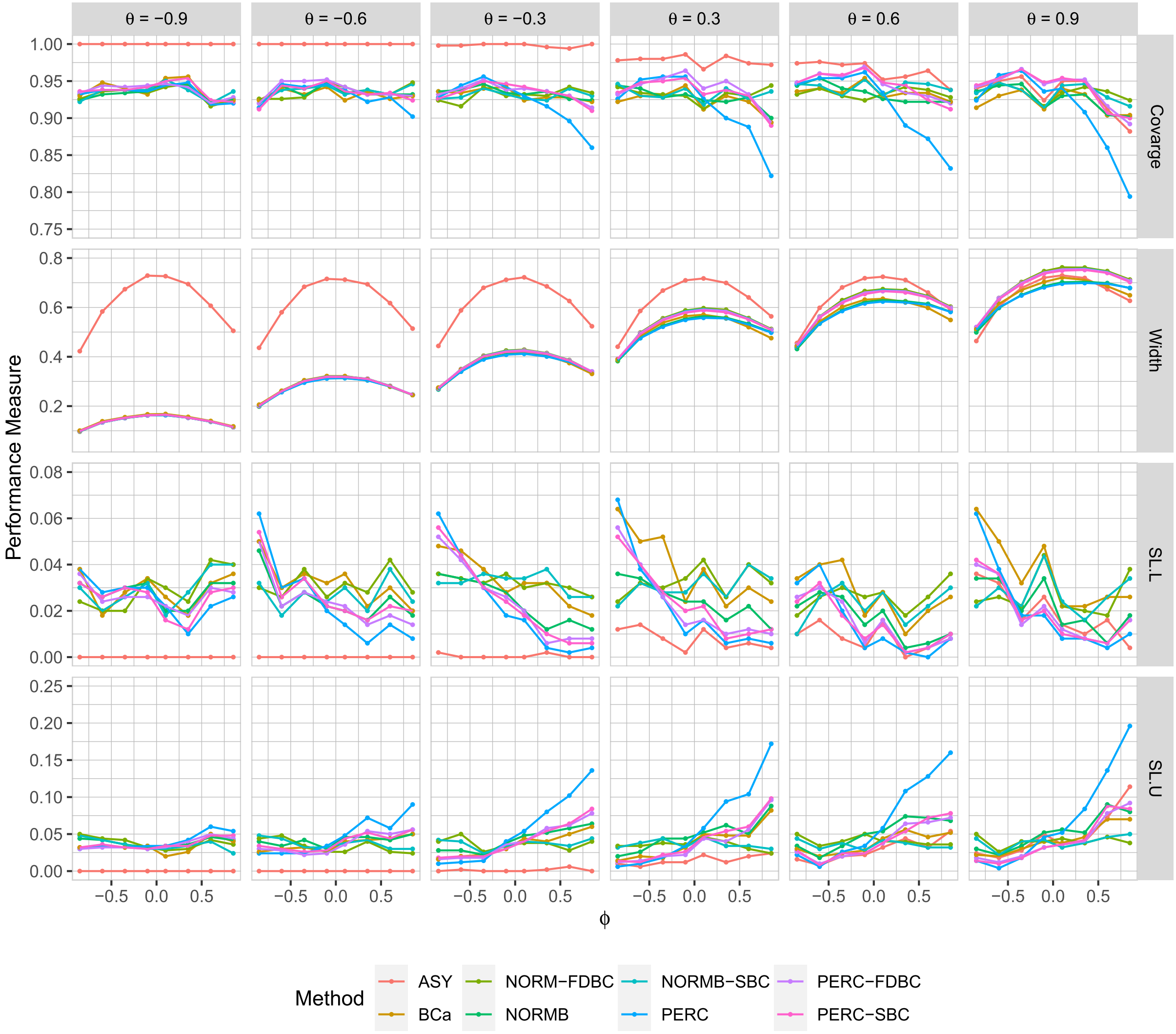

Figure 3: Performance of eight confidence interval procedures of ϕ_1 based on tree size $n = 63$ under the bivariate normal distribution for the errors ($\sigma = 0.5$).

$n = 63$ case. The results for the $n = 31$ and $n = 127$ cases can be found in the Supplementary Material.

We summarize the results in the following points:

- Considering the asymptotic confidence interval procedure (ASY) based on the plug-in estimate of the asymptotic standard error of $\hat{\phi}_1^{\mathrm{LS}}$, it is readily seen that this procedure has unsatisfactory performance as demonstrated by its excessively high coverage rates (near 100% in most cases) and overly large width, especially for negative θ values, e.g., $\theta = -0.9, -0.6$ or -0.3. The large width of the asymptotic confidence interval procedure can be attributed to the high variability in the plug-in estimate of the standard error of $\hat{\phi}_1^{\mathrm{LS}}$. As we shift towards positive values of θ, the relative performance of the asymptotic confidence interval procedure improves except for large positive ϕ_1 values, e.g., $\phi_1 = 0.9$, where the bias of LS estimator is known to be high. This confidence interval procedure does not seem to enjoy good symmetry

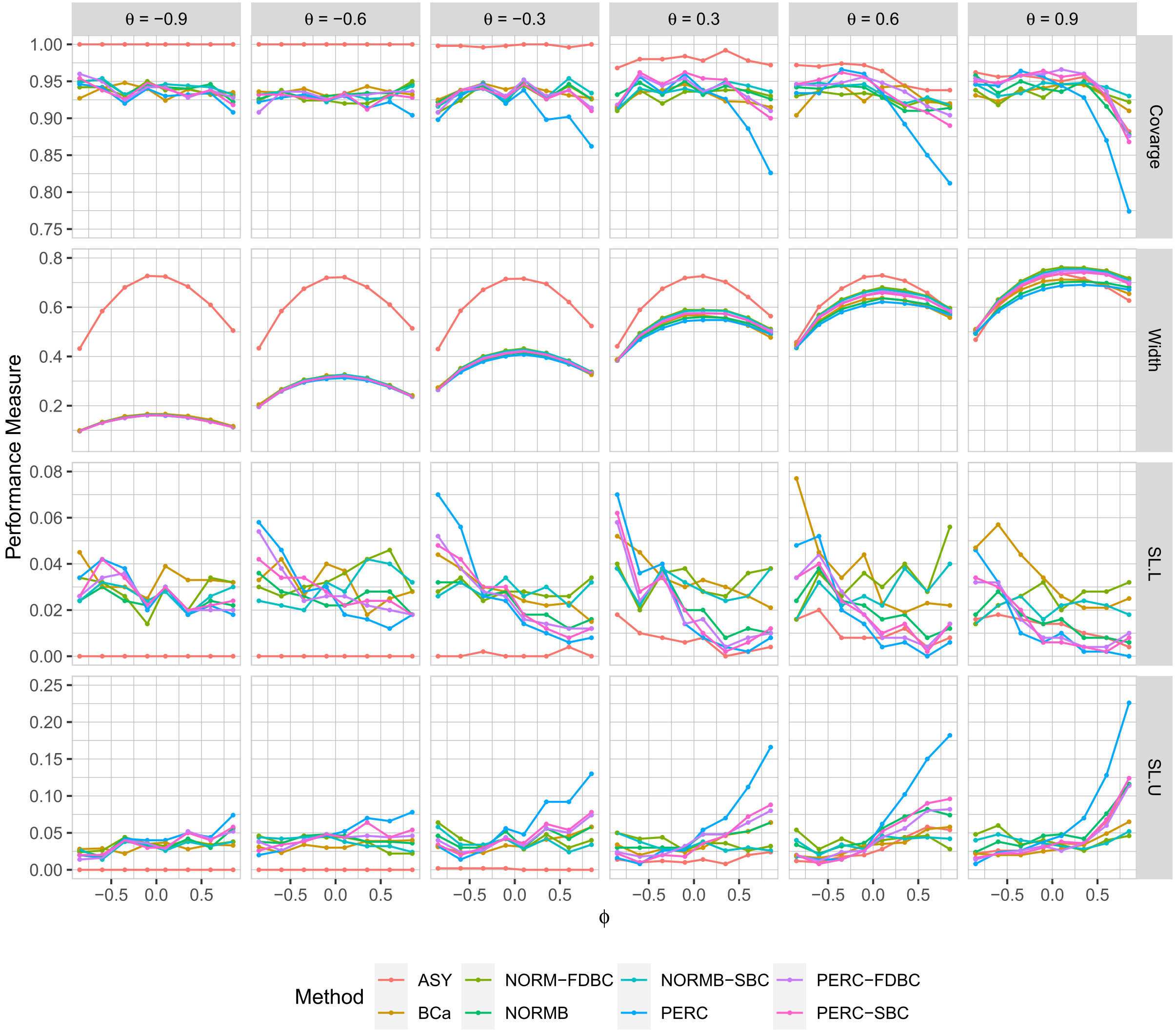

Figure 4: Performance of eight confidence interval procedures of ϕ_1 based on tree size $n = 63$ under the bivariate normal distribution for the errors ($\sigma = 1$).

levels as shown by the unbalanced significance rates (i.e., SL.L and SL.U).

- It is noteworthy that the uncorrected standard normal bootstrap (NORMB) confidence interval procedure has the best performance among the three uncorrected procedures; namely, ASY, NORMB, and the percentile bootstrap (PERC). Overall, the NORMB procedure enjoys reasonable coverage rates, width, and symmetry in most cases. On the contrary, the PERC procedure suffers from a dramatic drop in the coverage for large positive values of ϕ_1 and produces the most asymmetrical confidence intervals among all eight procedures. The PERC procedure enjoys good performance when both ϕ_1 and θ are negative, the combination where the bias of the LS estimate $\hat{\phi}_1^{\text{LS}}$ reaches its minimal level.
- The simulation results indicate that the bias-correction of the LS estimator significantly enhances the performance of the two bootstrap confidence interval procedures, NORMB and PERC. This is clear from the improved coverage, width, and symmetry of the NORMB-SBC,

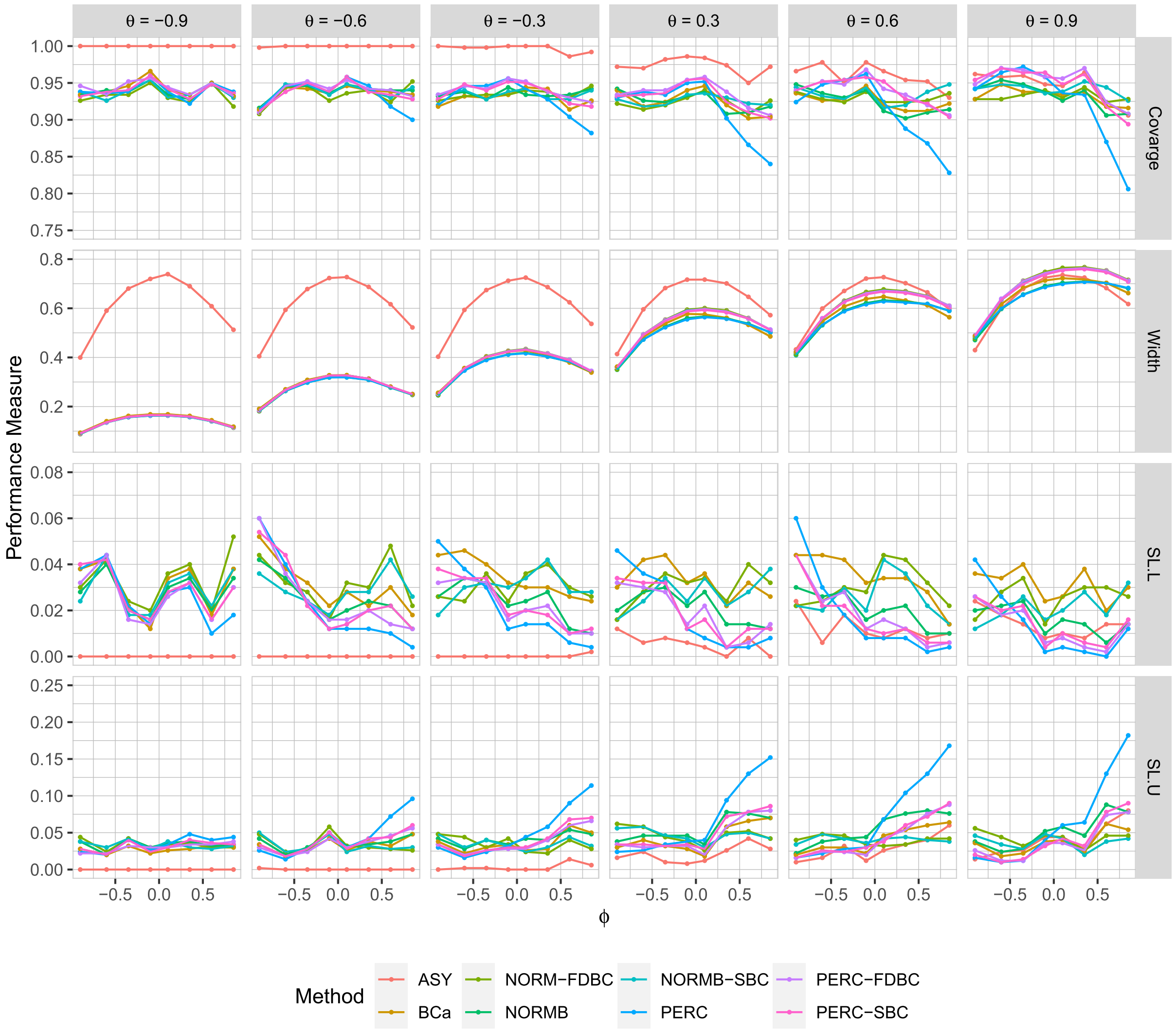

Figure 5: Performance of eight confidence interval procedures of ϕ_1 based on tree size $n = 63$ under the bivariate t-distribution with 10 degrees of freedom.

NORMB-FDBC, PERC-SBC, and PERC-FDBC procedures. The bias-corrected versions of the NORMB and PERC procedures perform quite similarly for negative values of θ with somewhat superior coverage for the corrected NORMB procedures at the expense of a slightly higher width. When θ is positive, the corrected PERC procedures have higher coverage rates, especially for $\phi_1 < 0.5$, and a smaller width than the corrected NORMB procedures. However, the corrected NORMB procedures appear to enjoy better symmetry than the corrected PERC procedures. In general, the two types of bias correction, namely, single bootstrap and fast-double bootstrap, have very similar performance. Considering the computational cost, the single bootstrap bias correction procedure is recommended over the fast-double bootstrap procedure.

- The BCa procedure has stable coverage rates across the different scenarios and produces confidence intervals that have a nice balance between coverage and width. Hence, the BCa

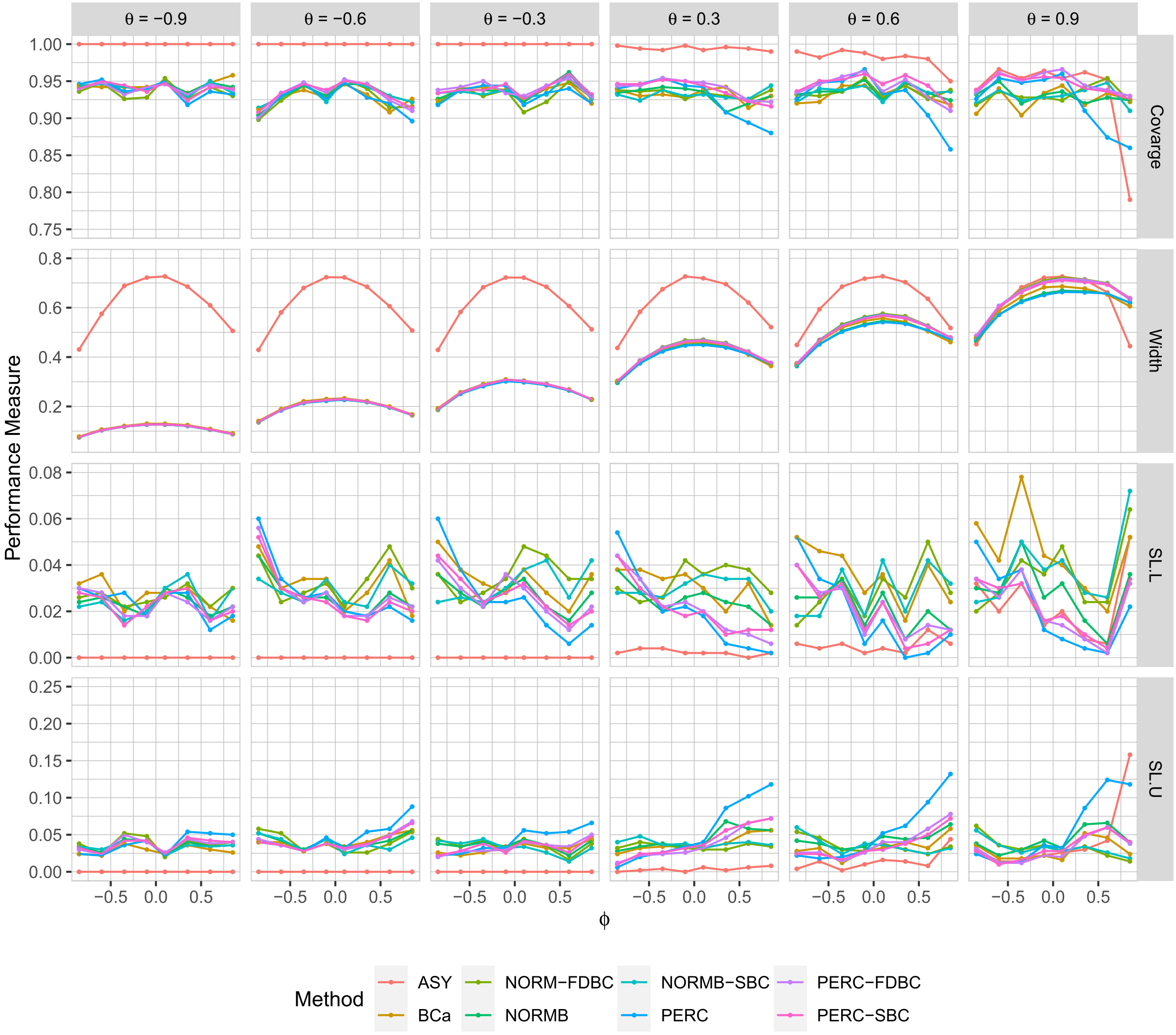

Figure 6: Performance of eight confidence interval procedures of ϕ_1 based on tree size $n = 63$ under the bivariate skew normal distribution with skewness parameter $a = 3$.

successfully addresses two of the main issues in the PERC bootstrap confidence interval procedure. However, the BCa does not fully treat the asymmetry in the PERC confidence intervals.

- We note that the observations summarized in the previous points hold regardless of the sample size and error distribution.
- The simulation results shown in the above figures and the figures in Supplementary Material suggest that the error distribution may impact the behavior of the confidence interval procedures analyzed in our study. However, the effect of the error distribution appears to be rather small except for the ASY procedure where we notice a dramatic drop in its performance for skew error distribution, especially when both θ and ϕ_1 are above 0.5.
- Lastly, upon inspection of the simulation results across the three sample sizes, $n = 31, 63, 127$, we notice that all eight confidence interval procedures are consistent. That is, at a fixed

combination of θ and ϕ_1, increasing the sample size leads to higher observed coverage rates and narrower intervals.

4.2 Applications to Cell Lineage Data

In this section, we apply the confidence interval procedures discussed in earlier sections to two real cell lineage datasets that can be modeled as bifurcating trees.

4.2.1 Lifetimes of EMT6 Cells

The lifetimes (in tenths of hours) of the progeny of EMT6 (BALB/c mouse mammary tumor) cells were recorded and 877 observations were obtained from 41 trees. The data were collected at the Institute of Cancer Research, Lille, France, and can be found in Appendix B of Staudte et al. (1984). The initial cells of all trees were missing and many of the trees had censored data resulting in non-perfect binary trees. Staudte et al. (1984) analyzed the data from all trees and noted that the mean lifetime was similar across trees. With the exception of two trees, all trees consisted of 63 or fewer cells. In this application, we combine observations from all 41 trees (after trimming the largest two trees at 63 cells) to build a perfect binary tree of 64 cells that can be modeled using the BAR model. To do so, we averaged observations from cells in each position in all 41 trees. Due to the varying trees sizes, the number of observations used in computing the cell averages also varied from one cell to another. Since the first cell was missing in all trees, we used the average of all cells' averages to estimate the initial cell for the averages tree. Figure 7 displays the resulting tree which is composed of five generations and 63

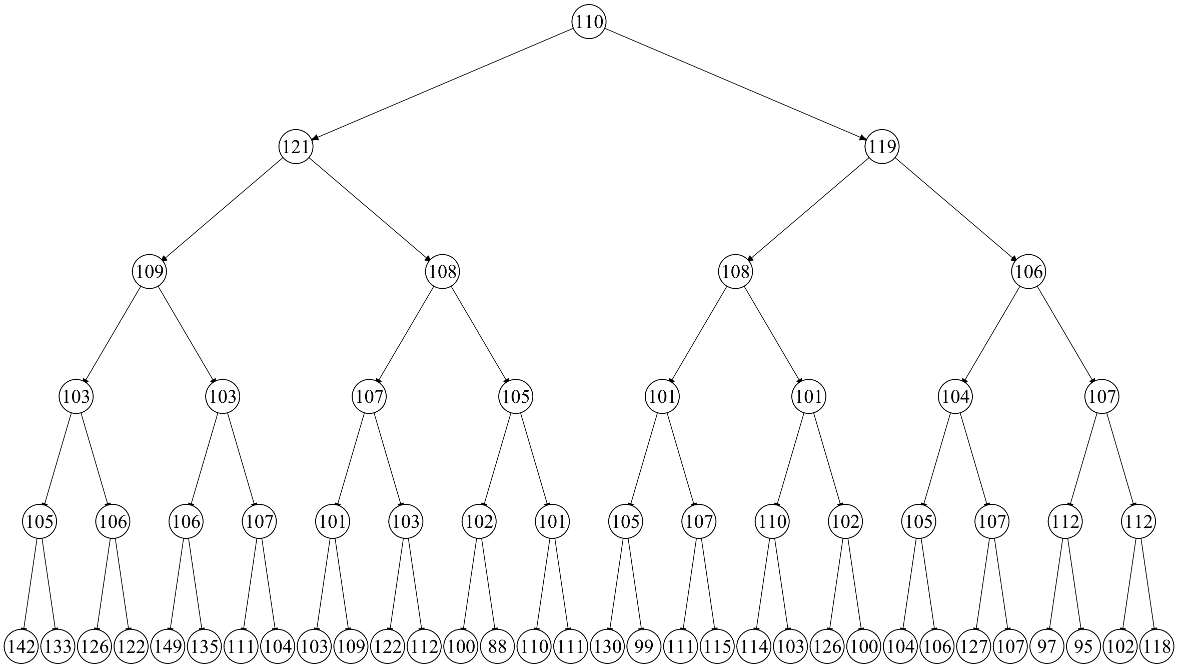

Figure 7: Average lifetimes (in tenths of hours) of EMT6 cells obtained from averaging observations from 41 bifurcating trees. The averaging of observations and the tree are made by the authors using the data in Staudte et al. (1984).

Table 1: 95% confidence intervals for the autoregressive coefficient ϕ_1 and their corresponding widths from the EMT6 and E. coli datasets.

Method	EMT6 ($\hat{\phi}_1 = -0.096, \hat{\theta} = 0.569$)		E. coli ($\hat{\phi}_1 = 0.355, \hat{\theta} = 0.584$)	
	95% CI	Width	95% CI	Width
ASY	$(-0.706, 0.513)$	1.219	$(-0.165, 0.876)$	1.041
NORMB	$(-0.438, 0.246)$	0.684	$(-0.168, 0.878)$	1.046
NORMB-SBC	$(-0.418, 0.236)$	0.654	$(-0.114, 1.000$[§]$)$	1.147
NORMB-FDBC	$(-0.378, 0.299)$	0.678	$(-0.169, 1.000$[§]$)$	1.205
PERC	$(-0.433, 0.197)$	0.630	$(-0.372, 0.708)$	1.081
PERC-SBC	$(-0.430, 0.278)$	0.708	$(-0.276, 0.866)$	1.142
PERC-FDBC	$(-0.367, 0.296)$	0.663	$(-0.249, 0.913)$	1.162
BCa	$(-0.328, 0.326)$	0.654	$(0.039, 1.000$[§]$)$	1.070

[§] The upper limit for the confidence interval was trimmed at 1.000 as ϕ_1 is restricted to be between 0 and 1 by the stationarity assumption. The calculation of the confidence interval width uses the upper limit before trimming.

observations. The LS estimates for the first-order BAR model coefficients from this tree were found to be $\hat{\phi}_0 = 120.25$ and $\hat{\phi}_1 = -0.096$ with an estimated errors correlation of $\hat{\theta} = 0.569$. We also computed 95% confidence intervals for the autoregressive coefficient ϕ_1 using the eight confidence interval procedures presented earlier. These confidence intervals and their widths are shown in Table 1. We note that all eight confidence intervals contain the value zero suggesting that the true mother-daughter correlation, ϕ_1, is not significantly different from zero. The Wald-Type confidence interval (ASY) is much wider than all other intervals which is consistent with what was observed from the simulations. Assuming that the true value of ϕ_1 is zero, the BCa confidence interval is the only procedure that produces a symmetric confidence interval with reasonable width.

4.2.2 Lifetimes of E. Coli Cells

In this application, we use data on the lineage of E. coli cells. The data is taken from Cowan and Staudte (1986). Observations in the data represent the lifetimes (in minutes) of the E. coli cells. The 31 observations form a perfect binary tree with four generations as displayed in Figure 1. The LS estimates of the first-order BAR model coefficients from this data are $\hat{\phi}_0 = 17.617$ and $\hat{\phi}_1 = 0.355$ with errors correlation $\hat{\theta} = 0.584$. In the last two columns of Table 1, we report the results of 95% confidence intervals for the autoregressive coefficient ϕ_1 calculated using the eight confidence interval procedures studied in this paper. Although the LS estimate of ϕ_1 is 0.355, which is much larger than 0, seven of the eight confidence intervals contain the value zero. For the uncorrected procedures, i.e., ASY, NORMB, and PERC, this behavior is likely due to the large amount of negative bias in the LS estimate which is reported by Elbayoumi and Mostafa (2021b) to be about 25%. The increase in width of the four bias-corrected confidence interval procedures, NORMB-SBC, NORMB-FDBC, PERC-SBC and PERC-FDBC, can be attributed to the fact that the bias-correction can lead to a significant increase in the variance, especially for small sample sizes such as $n = 31$ (see, Elbayoumi and Mostafa, 2021b). Similar to the EMT6 cells application above, the BCa procedure seems to produce a more accurate inference for ϕ_1 as indicated by the all-positive interval while maintaining a reasonable width.

5 Discussions

In this paper, we studied the effect of the bias in the least squares estimation of the autoregressive coefficients in bifurcating autoregressive (BAR) models on the performance of different types of bootstrap confidence intervals. Specifically, we focused on constructing and evaluating bias-corrected/uncorrected bootstrap confidence intervals for the autoregressive coefficient ϕ_1 in the BAR(1) model. Both single and fast-double bootstrap bias corrections were used in the bias-corrected bootstrap confidence intervals. The behavior of the uncorrected and corrected confidence interval procedures was examined through extensive simulations and two cell lineage data sets. Several concluding points can be drawn from the simulation results and real data examples. First, the performance of uncorrected bootstrap confidence interval procedures, namely; standard normal and percentile bootstrap confidence intervals, can be significantly affected by the bias in the LS estimator. Second, the standard normal bootstrap confidence interval is generally superior to the percentile bootstrap procedure. Third, the correction of the bias in the LS estimator significantly enhances the performance of these two bootstrap confidence interval procedures with the bias-corrected standard normal bootstrap procedure having better coverage rate, especially when the errors correlation is negative, and better symmetry in most cases. Fourth, the two types of bias correction –single bootstrap and fast-double bootstrap– have similar performance and, therefore, the single bootstrap bias correction procedure is recommended due to its computational efficiency relative to the fast-double bootstrap. Finally, although the bias-corrected and accelerated (BCa) confidence interval procedure does not fully address the asymmetry problem in the percentile bootstrap confidence interval, it produces reasonably tight confidence intervals that have stable coverage rates. This preferable performance of the BCa procedure is further confirmed in the two real data examples.

It is worth noting that the work presented in this paper can be extended to higher-order BAR models which can be quite useful when, in addition to the correlations between the immediate relatives in the tree, the researcher is interested in the correlations between distant relatives such as cousins (Huggins and Basawa, 1999). Elbayoumi and Mostafa (2021b) discussed the bias in the LS estimation of higher-order BAR models, especially BAR(2) models. They showed that the LS estimators of the BAR(2) model coefficients have similar biases as in the case of BAR(1) models and outlined the extension of bootstrap bias correction methods to BAR(2) models. It is straightforward to extend the bootstrap confidence interval procedures discussed above for the case of BAR(1) models to construct confidence intervals for the coefficients of BAR(p) models. This extension is straightforward since the bootstrap bias correction and bootstrap confidence intervals discussed here use model-based bootstrapping which makes the necessary modifications for higher-order BAR models quite obvious. The finite sample behavior of bias-corrected/uncorrected bootstrap confidence intervals for BAR(p), $p > 1$, models can be investigated via an empirical study similar to the one reported in this paper.

Supplementary Material

The supplementary material includes the following files and folders: (1) README: a brief explanation of all the files and folders in the supplementary material; (2) The application datasets; (3) Code files; and (4) Additional simulation results.

Acknowledgement

The authors are indebted to the Editor, Dr. Jun Yan, for his comments and suggestions that led to significant enhancements in the final version of the paper.

References

Berkowitz J, Kilian L (2000). Recent developments in bootstrapping time series. *Econometric Reviews*, 19(1): 1–48. https://doi.org/10.1080/07474930008800457

Bui Q, Huggins R (1999). Inference for the random coefficients bifurcating autoregressive model for cell lineage studies. *Journal of Statistical Planning and Inference*, 81(2): 253–262. https://doi.org/10.1016/S0378-3758(99)00049-X

Chang J, Hall P (2015). Double-bootstrap methods that use a single double-bootstrap simulation. *Biometrika*, 102: 203–214. https://doi.org/10.1093/biomet/asu060

Cowan R (1984). Statistical concepts in the analysis of cell lineage data. In: *1983 Workshop Cell Growth Division*, 18–22. Latrobe University, Melbourne.

Cowan R, Staudte RG (1986). The bifurcating autoregression model in cell lineage studies. *Biometrics*, 42: 769–783. https://doi.org/10.2307/2530692

Efron B (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82: 171–200. https://doi.org/10.1080/01621459.1987.10478410

Elbayoumi T, Mostafa S (2021a). *bifurcatingr: Bifurcating Autoregressive Models*. R package version 1.0.0.

Elbayoumi T, Terpstra J (2016). Weighted L1-estimates for the first-order bifurcating autoregressive model. *Communications in Statistics. Simulation and Computation*, 45: 2991–3013. https://doi.org/10.1080/03610918.2014.938826

Elbayoumi TM, Mostafa SA (2021b). On the estimation bias in first-order bifurcating autoregressive models. *Stat*, 10(1): e342.

Hall P (1992). *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, New York.

Hawkins ED, Markham JF, McGuinness LP, Hodgkin PD (2009). A single-cell pedigree analysis of alternative stochastic lymphocyte fates. *Proceedings of the National Academy of Sciences*, 106(32): 13457–13462. https://doi.org/10.1073/pnas.0905629106

Huggins R, Basawa I (2000). Inference for the extended bifurcating autoregressive model for cell lineage studies. *Australian & New Zealand Journal of Statistics*, 42: 423–432. https://doi.org/10.1111/1467-842X.00139

Huggins RM (1995). A law of large numbers for the bifurcating autoregressive process. *Communications in Statistics–Stochastic Models*, 11(2): 273–278. https://doi.org/10.1080/15326349508807345

Huggins RM, Basawa IV (1999). Extensions of the bifurcating autoregressive model for cell lineage studies. *Journal of Applied Probability*, 36: 1225–1233. https://doi.org/10.1239/jap/1032374768

Kimmel M, Axelrod D (2005). *Branching Processes in Biology*. Springer-Verlag, New York.

Lee SMS, Young GA (1999). The effect of Monte Carlo approximation on coverage error of double-bootstrap confidence intervals. *Journal of the Royal Statistical Society: Series B*, 61(2): 353–366. https://doi.org/10.1111/1467-9868.00181

Liu-Evans GD, Phillips GD (2012). Bootstrap, Jackknife and COLS: Bias and mean squared error in estimation of autoregressive models. *Journal of Time Series Econometrics*, 4(2): 1–33.

https://doi.org/10.1515/1941-1928.1122

Ouysse R (2013). A fast iterated bootstrap procedure for approximating the small-sample bias. *Communications in Statistics – Simulation and Computation*, 42(7): 1472–1494. https://doi.org/10.1080/03610918.2012.667473

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sandler O, Mizrahi SP, Weiss N, Agam O, Simon I, Balaban NQ (2015). Lineage correlations of single cell division time as a probe of cell-cycle dynamics. *Nature*, 519: 468–471. https://doi.org/10.1038/nature14318

Shi SG (1992). Accurate and efficient double-bootstrap confidence limit method. *Computational Statistics and Data Analysis*, 13: 21–32. https://doi.org/10.1016/0167-9473(92)90151-5

Staudte R, Guiguet M, d'Hooghe MC (1984). Additive models for dependent cell populations. *Journal of Theoretical Biology*, 109(1): 127–146. https://doi.org/10.1016/S0022-5193(84)80115-0

Terpstra JT, Elbayoumi T (2012). A law of large numbers result for a bifurcating process with an infinite moving average representation. *Statistics & Probability Letters*, 82(1): 123–129. https://doi.org/10.1016/j.spl.2011.09.012

Wuertz D, Setz T, Chalabi Y, Smith P (2022). *fCopulae: Rmetrics – Bivariate Dependence Structures with Copulae*. R package version 4021.84.

Zhou J, Basawa I (2005). Least-squares estimation for bifurcating autoregressive processes. *Statistics & Probability Letters*, 74(1): 77–88.