On the Estimation Bias in First-Order Bifurcating Autoregressive Models

Tamer Elbayoumi
Assistant Professor
Department of Mathematics and Statistics
North Carolina A&T State University

Sayed Mostafa
Assistant Professor
Department of Mathematics and Statistics
North Carolina A&T State University

Points will talk about

- The Bifurcating Autoregressive (BAR) Model
- The Least Squares Estimation (LSE) for the BAR model
- The Problem and the Goal
- The Bias in the LS estimators
- The Bootstrap Bias Correction Methods
- Simulation Results
- The Asymptotic Bias Correction Formula
- In Progress

Bifurcating Autoregressive (BAR) Model

- Bifurcating Autoregressive (BAR) Model is an adaptation of autoregressive (AR) model to binary tree structured data, where each individual observation in any generation gives rise to two offspring in the next generation.
- First introduced by Cowan and Staudte (1986).
- Modeling Cell Lineage Data in Biology.
- This model allows to the sister cells be correlated.
- All cells are correlated.

The Bifurcating Autoregressive BAR (*p*) Model

The *p*th-order bifurcating autoregressive process (BAR(*p*)) is defined by the equation

$$X_t = \phi_0 + \phi_1 X_{\left\lfloor \frac{t}{2} \right\rfloor} + \phi_2 X_{\left\lfloor \frac{t}{4} \right\rfloor} + \cdots + \phi_p X_{\left\lfloor \frac{t}{2^p} \right\rfloor} + \varepsilon_t, \qquad \text{for all t} \geq 2^p\ , p \geq 1$$

where

- X_t is the observations on a perfect binary tree with g generations.
- $X_{\left\lfloor \frac{t}{2^p} \right\rfloor}$ are the ancestors of X_t, where $[u]$ defines the largest integer $\leq u$.
- $\phi_0, \phi_1, \phi_2, \ldots, \phi_p$ are the parameters that need to be estimated.

The Bifurcating Autoregressive BAR (*1*) Model

- $\{(\varepsilon_{2t}, \varepsilon_{2t+1}),\ \mathrm{t} \geq 1\}$ are independently and identically distributed (iid) bivariate random variables with mean zero and a variance–covariance structure,

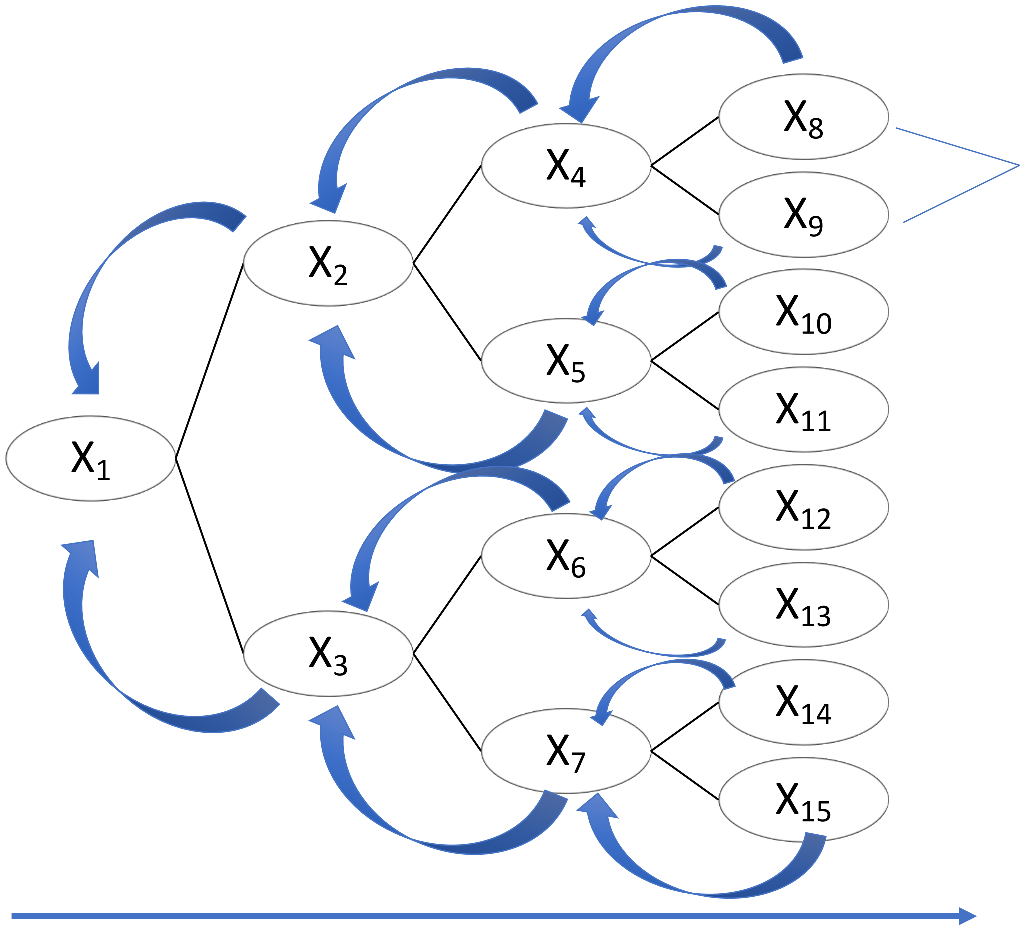
$$\begin{pmatrix} 1 & \theta \\ \theta & 1 \end{pmatrix} \sigma^2,$$

where

- θ is the correlation between $(\varepsilon_{2t}, \varepsilon_{2t+1})$ or the environmental effect.
- σ^2 is the variance of ε_{2t} and ε_{2t+1}.
- It is assumed that $\phi_t \in (-1, 1), t = 1, \ldots, p$. This implies that X_t is causal and invertible and it implies that the process is stationary.
- The correlation coefficient between the sisters (X_{2t}, X_{2t+1}) is defined as ρ and it is given by

$$\rho = \phi_1^2 + (1 - \phi_1^2)\theta.$$

The Bifurcating Autoregressive BAR (1) Model



Sisters are correlated

Let $X_1, X_2, \dots, X_n$ denote the random variables corresponding to observations on a perfect binary tree with g generations. The initial observation X_1 corresponds to generation 0, while the observations $X_{2^i}, X_{2^i+1}, \dots, X_{2^{i+1}-1}$ correspond to the 2^i observations in generation $i = 1,2, \dots, g$. Note that $n = 2^{g+1} - 1$.

The Bifurcating Autoregressive BAR (1) Model

The BAR(1) model is given by

$$X_t = \phi_0 + \phi_1 X_{\left[\frac{t}{2}\right]} + \varepsilon_t, \qquad \text{for all t} \geq 2$$

where

- X_t is an observed value at time t.
- $X_{\left[\frac{t}{2}\right]}$ is the mother of X_t for all $t \ \geq 1$, where $[u]$ defines the largest integer ≤ u.
- ϕ_0 and ϕ_1 are the parameters that need to be estimated, where ϕ_1 denotes the maternal correlation or the inherited effect.

Previous Works Summary

- Previous works concentrated on
  - Modeling under stationary assumption. (As it is assumed in this study).
  - Modeling under non-stationary assumption.
  - Variability between trees.
  - Asymptotic distribution of the estimated parameters.
  - The Law of Large Numbers.
- Estimation methods are used
  - MLE under normality.
  - Modified MLE to deal with outliers.
  - Nonparametric estimation to deal with outliers.
  - Least Squares Estimation. (It is used in this study)

Least Squares Estimation

In 2005, Zhou and Basawa introduced the least squares (LS) estimators for the BAR model and derived the asymptotic distribution for the estimators. The LS estimators of BAR(1) are given by

$$\hat{\phi}_1 = \frac{\sum_{t=1}^{m} X_t(U_t - \bar{U}_t)}{\sum_{t=1}^{m}(X_t - \bar{X})^2}, \hat{\phi}_0 = \bar{U}_t - \hat{\phi}_1\bar{X},$$

$$U_t = \frac{X_{2t}+X_{2t+1}}{2}, \bar{X} = \frac{1}{m}\sum_{t=1}^{m} X_t\,, \bar{U}_t = \frac{1}{m}\sum_{t=1}^{m} U_t\,, m = \frac{n-1}{2},$$

and m the number of triplets (X_t, X_{2t}, X_{2t+1}).

Least Squares Estimation

When substituting U_t in the $\hat{\phi}_1$ and $\hat{\phi}_0$ equations, it gives the common estimated equations of LS

$$\hat{\phi}_1 = \frac{\sum_{t=1}^{n} X_t \left(X_{\left[\frac{t}{2}\right]} - \bar{X}_t \right)}{\sum_{t=1}^{n} (X_t - \bar{X}_t)^2}$$

$$\hat{\phi}_0 = \left(1 - \hat{\phi}_1\right)\bar{X}_t$$

Where $\bar{X}_t = \frac{1}{n} \sum_{t=1}^{n} \bar{X}_t$.

Least Squares Estimation

The asymptotic distribution for the LS of BAR(1) estimators are given by

$$\sqrt{n}\left(\hat{\phi}-\phi\right) \overset{\mathcal{D}}{\to} \mathcal{N}\left(0, \sigma^2(1+\theta)A^{-1}\right),$$

where

$$A = \begin{pmatrix} 1 & \phi_0/(1-\phi_1) \\ \phi_0/(1-\phi_1) & \dfrac{\sigma^2}{1-\phi_1^2} + \left(\dfrac{\phi_0}{(1-\phi_1)}\right)^2 \end{pmatrix},$$

ϕ denotes the vector of the true parameters containing ϕ_0 and ϕ_1,

The Problem

- It is well-known that the Least Square estimators of autoregressive (AR) models are biased in small samples.
- The mean-bias of the LSE of AR is of order $O(n^{-1})$*.
- The LS estimation is widely used in modeling the cell division process.
- Investigation regarding the bias of LSE for BAR models is needed.

The Goal

- Studying the bias of LS estimates for BAR(1).
- Proposed a bootstrap approach for correcting the bias in the LS estimates.
- Proposed an inference based on the Confidence Intervals for the adjusted estimates.

* Marriott and Pope (1954), and Kendall (1954).

The Bias of LSE for BAR(1)

Our primary interest is to study the behavior of ϕ_1 parameter via Monte Carlo. 10000 realizations of perfect binary tree sized 31,63,127 and 255 are generated, based on combinations of ϕ_1= -0.75,-0.5,-0.25,0.25,0.5, and 0.75, and θ = -0.8,-0.4,-0.2,0.2,0.4, and 0.8. The intercept ϕ_0 is set to 10 for all simulations. For all generated binary trees, it is assumed stationary and the initial observation X_1 is randomly selected from a simulated large binary tree of size 127.

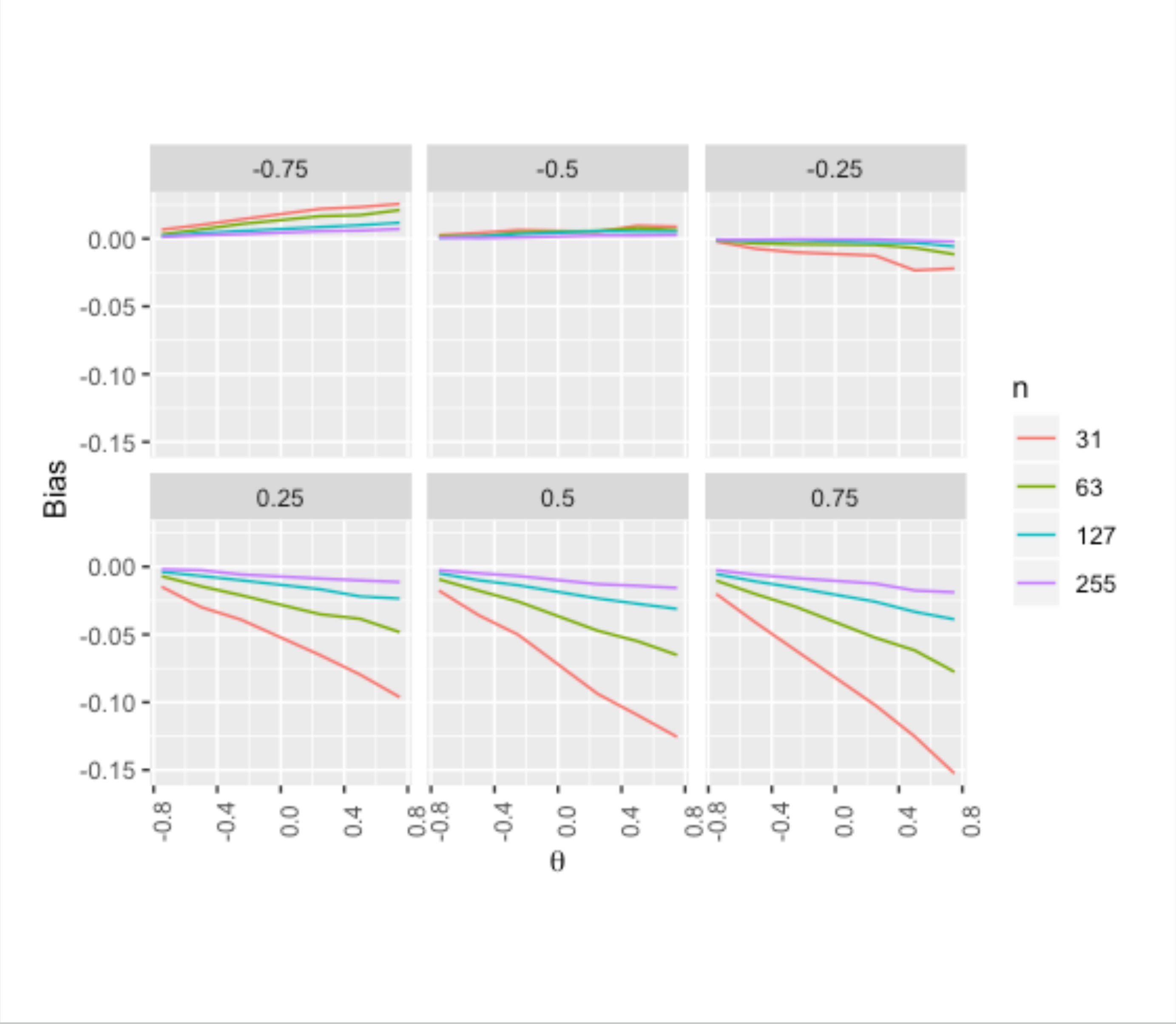
The stationary assumption implies that the bias does not depend on the variance σ^2.

The empirical bias is calculated as

$$Empirical\ Bias(\hat{\phi}_1) = \frac{1}{nsim}\sum_{i=1}^{nsim}(\hat{\phi}_{boot,i} - \hat{\phi}_{LS,i}).$$

The Bias of LSE for BAR(1)

- The LSE bias is a linear function in ϕ_1 and θ.
- The bias appears to be of order $O(n^{-1})$.
- For positive values of ϕ_1 and close to zero, the underestimated LSEs are found.
- The bias increases as the value of θ increases from -1 to +1.
- For negative values of ϕ_1 from (-0.5) to -1, the overestimated LSEs are found.
- The bias decreases as the sample size increases.



The Bootstrap Bias Correction Methods

- **Single bootstrap bias correction**
- **Double bootstrap bias correction**
- **Fast double bootstrap bias correction**

The Bootstrap Bias Correction Methods

Single Bootstrap Bias Correction Algorithm

Given the original sample X_t^n, compute the LS estimates $\hat{\phi}_0$ and $\hat{\phi}_1$, and the estimated errors $\hat{e}_{2t}$ and $\hat{e}_{2t+1}$ for all $t \geq 1$. From the estimated errors, draw B bootstrap samples of size $(n-1)/2$ of the pairs $(\hat{e}_{2t}, \hat{e}_{2t+1})$ and form X_t^{n*}, $b = 1, \dots, B$ by

$$X_{2t}^{n*} = \hat{\phi}_0 + \hat{\phi}_1 X_t^n + \hat{e}_{2t}, \text{ and } X_{2t+1}^{n*} = \hat{\phi}_0 + \hat{\phi}_1 X_t^n + \hat{e}_{2t+1}$$

For all bootstrap binary tree samples, we keep the initial value $X_1^* = X_1$. Next, for each bootstrap binary tree sample compute the $\hat{\phi}_{1b}^*$, $b = 1, \dots, B$. Then, define the estimated bias $\beta_{\hat{\phi}_1}$ as,

$$\beta_{\hat{\phi}_1} = \frac{1}{B} \sum_{b=1}^{B} (\hat{\phi}_{1,b}^* - \hat{\phi}_1).$$

Finally, The adjusted LS estimate is $\hat{\phi}_1$- $\beta_{\hat{\phi}_1}$.

The Bootstrap Bias Correction Methods

Double Bootstrap Bias Correction Algorithm

Given a resampling data $X_b^{n*}, b = 1, \dots, B_1$ from the single bootstrap BAR(1) algorithm, define a second phase of resampling based on the LS estimates $\hat{\phi}_0^*, \hat{\phi}_1^*$, and the errors $(\hat{e}_{2t}^*, \hat{e}_{2t+1}^*)$ and form $X_t^{n**}, b = 1, \dots, B_2$,

$$X_{2t}^{n**} = \hat{\phi}_0^* + \hat{\phi}_1^* X_t^* + \hat{e}_{2t}^*, \text{ and } \; X_{2t+1}^{n**} = \hat{\phi}_0^* + \hat{\phi}_1^* X_t^* + \hat{e}_{2t+1}^*$$

For all second phase bootstrap binary tree samples, we keep the initial value $X_1^{**n} = X_1^{n*}$. Next, for each second phase resampling samples, compute the $\hat{\phi}_{1b}^{**}$, $b = 1, \dots, B_1$. Then, define the additive adjustment $\gamma_{\hat{\phi}_1}$ as,

$$\gamma_{\hat{\phi}_1} = \frac{1}{B_1} \sum_{b=1}^{B_1} \beta^*_{\hat{\phi}_{1,b}} - \beta_{\hat{\phi}_1}, \text{where} \quad \beta^*_{\hat{\phi}_1} = \hat{\phi}^*_{1,b} - \frac{1}{B_2} \sum_{j=1}^{B_2} \hat{\phi}^{**}_{1b,j}$$

Finally, The adjusted LS estimate is $\hat{\phi}_1 - \beta_{\hat{\phi}_1} - \gamma_{\hat{\phi}_1}$, where $\beta_{\hat{\phi}_1}$ is the estimated bias from the single bootstrap bias correction algorithm .

The Bootstrap Bias Correction Methods

Fast Double Bootstrap Bias Correction Algorithm*

This method is similar the double bootstrap bias correction method. The only difference is the resampling in the second phase for only one bootstrap sample instead of B_2 samples. The additive adjustment $\gamma_{\hat{\phi}_1}$ as,

$$\gamma_{\hat{\phi}_1} = \frac{1}{B_1}\sum_{b=1}^{B_1} \beta^{*}_{\hat{\phi}_1,b} - \beta_{\hat{\phi}_1}, \text{where} \quad \beta^{*}_{\hat{\phi}_1} = \hat{\phi}^{*}_{1,b} - \frac{1}{B_2}\sum_{j=1}^{B_2} \hat{\phi}^{**}_{1b,j}$$

Finally, The corrected LS estimate is $\hat{\phi}_1 - \beta_{\hat{\phi}_1} - \gamma_{\hat{\phi}_1}$, where $\beta_{\hat{\phi}_1}$ is the estimated bias from the single bootstrap bias correction algorithm .

* Ouysse (2013).

The Bootstrap Bias Correction Methods

10000 realizations of perfect binary tree sized 31,63, and 127 are generated, based on combinations of ϕ_1= -0.75,-0.5,-0.25,0.25,0.5, and 0.75 and θ = -0.8,-0.4,-0.2,0.2,0.4, and 0.8. The intercept ϕ_0 is set to 10 for all simulations. For all generated binary trees, it is assumed stationary and the initial observation X_1 is randomly selected from a simulated large binary tree of size 127.

The following graphs show the bias of the corrected LS estimates for the three bootstrap methods comparing to the original LS estimates:

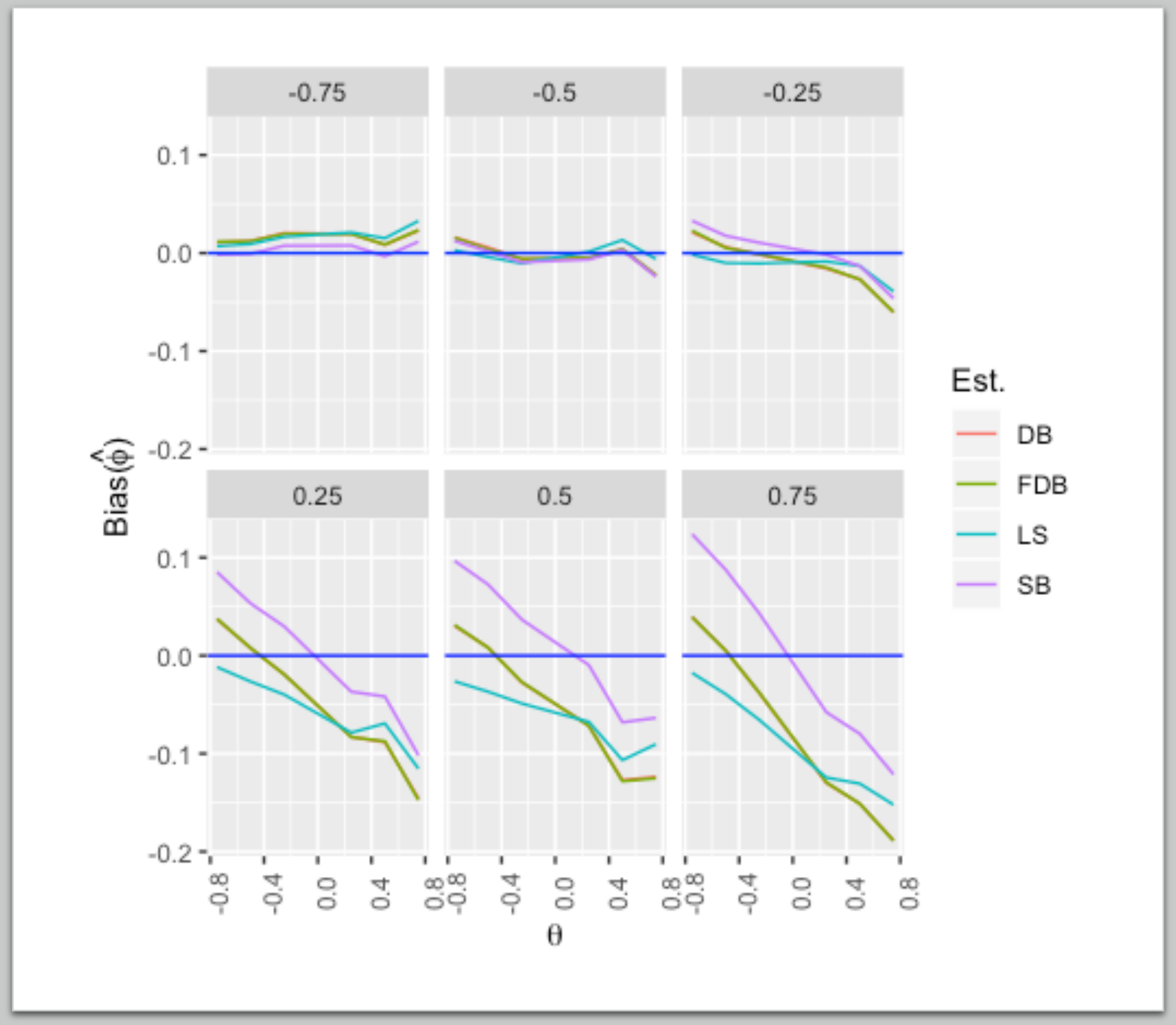
- Single bootstrap bias correction
- Double bootstrap bias correction
- Fast double bootstrap bias correction

The Bootstrap Bias Correction Results of Sample Size 31

- Double bootstrap and Fast double bootstrap methods failed to correct the bias in the LS estimate of the slope of the BAR(1).

- The single bootstrap corrected the bias in the LS estimate of the slope of the BAR(1).

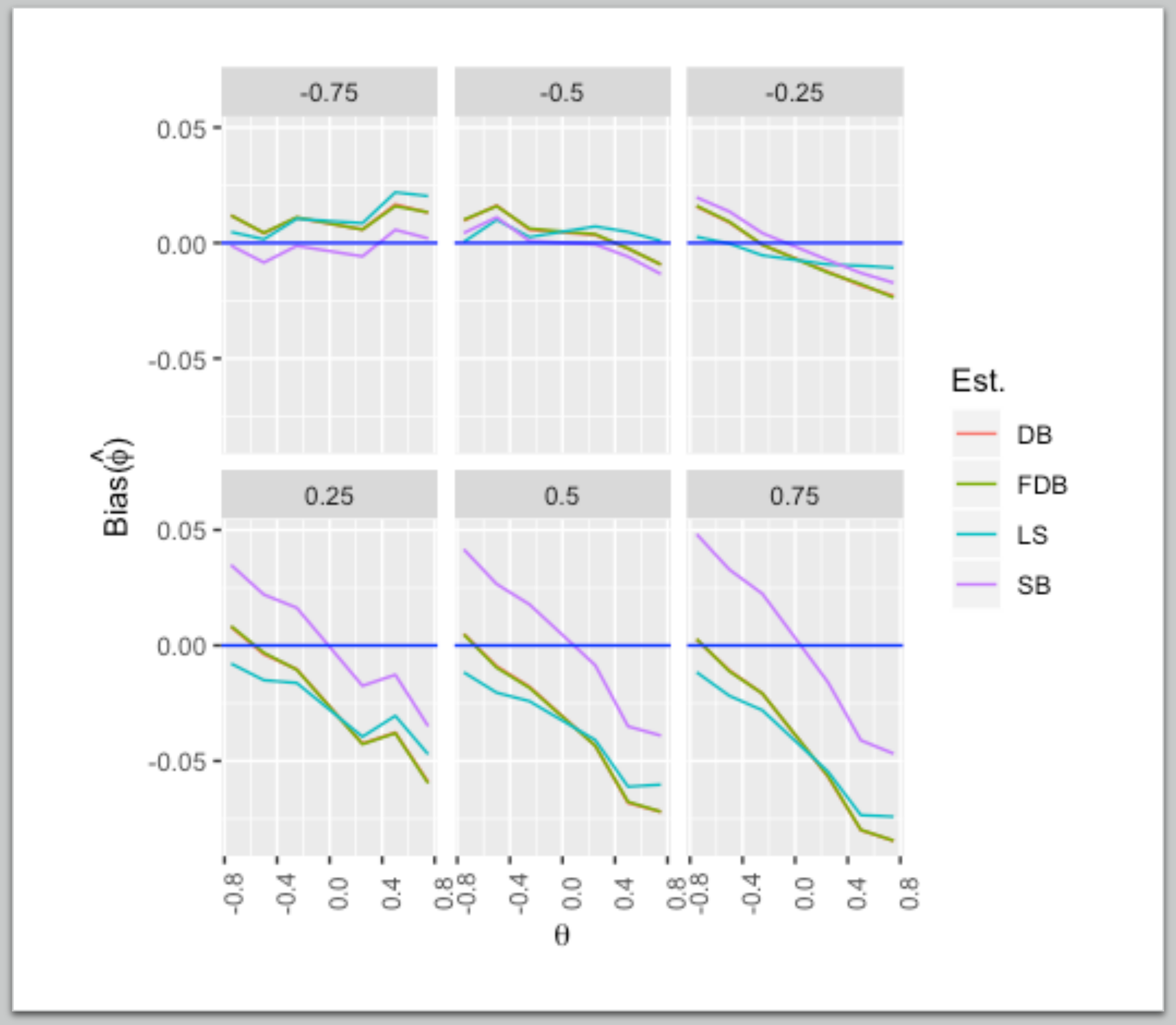
- The bias is too small when θ is zero or close to zero.



The Bootstrap Bias Correction Results of Sample Size 63

- Still some bias exists because of the well-known problem "persistent excursion" bias*.

This awkward problem arises when the ϕ_1 is positive and is emphasized further when θ is also positive. The problem is like one that occurs in the theory of autoregressive time series (though curiously ignored in that literature).

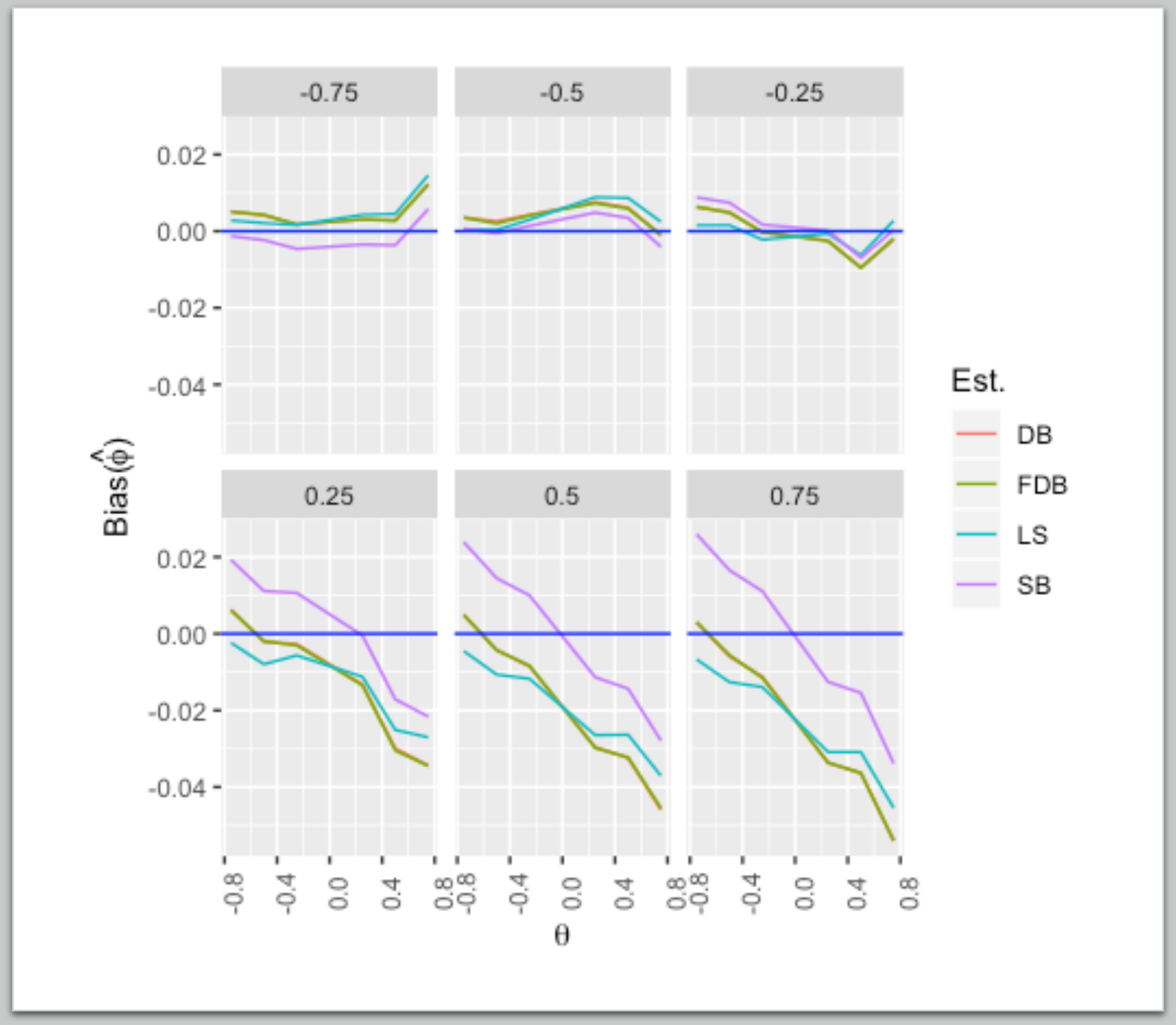


* Cowan and Staudte (1986).

The Bootstrap Bias Correction Results of Sample Size 127

- Increasing the sample size decreases the persistent excursion bias by half.

- This is consistent with the literature that discussed the bias of the estimators of autoregressive models.



The Asymptotic Bias Correction Formula

Theorem: *The asymptotic bias of the LS estimator for the slope of the BAR(1) model with an intercept is given by*

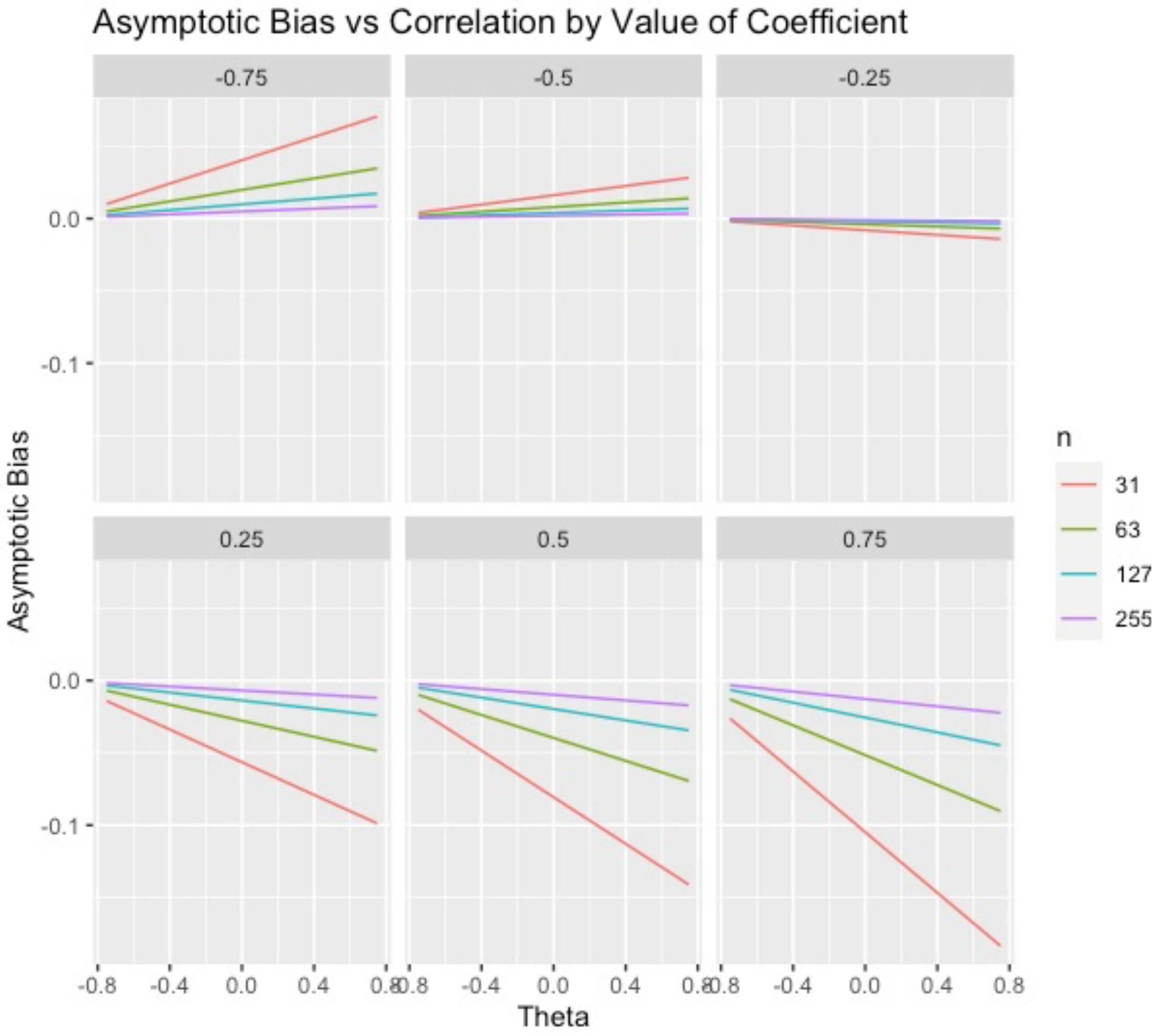
$$-\frac{1}{n}(1+\theta)(1+3\phi_1).$$

The proof:

The proof is too much similar to the Marriott and Pope (1954)'s proof, for deriving the asymptotic bias of the slope of first-order autoregressive model.

The Asymptotic Bias Correction Formula

ϕ_1	θ	n	LS Bias	Asy. Bias
0.75	-0.8	31	-0.0200	-0.0262
0.75	-0.4	31	-0.0414	-0.0524
0.75	-0.2	31	-0.0617	-0.0786
0.75	0.2	31	-0.1021	-0.1310
0.75	0.4	31	-0.1253	-0.1573
0.75	0.8	31	-0.1525	-0.1835
0.75	-0.8	63	-0.0103	-0.0129
0.75	-0.4	63	-0.0203	-0.0258
0.75	-0.2	63	-0.0295	-0.0387
0.75	0.2	63	-0.0523	-0.0645
0.75	0.4	63	-0.0617	-0.0774
0.75	0.8	63	-0.0775	-0.0903
0.75	-0.8	127	-0.0055	-0.0064
0.75	-0.4	127	-0.0109	-0.0128
0.75	-0.2	127	-0.0155	-0.0192
0.75	0.2	127	-0.0258	-0.0320
0.75	0.4	127	-0.0333	-0.0384
0.75	0.8	127	-0.0388	-0.0448
0.75	-0.8	255	-0.0026	-0.0032
0.75	-0.4	255	-0.0059	-0.0064
0.75	-0.2	255	-0.0084	-0.0096
0.75	0.2	255	-0.0124	-0.0159
0.75	0.4	255	-0.0175	-0.0191
0.75	0.8	255	-0.0189	-0.0223



Conclusion

- For small samples of the Binary trees, the bias in the least-squares estimators for the BAR model exists and has an order of $O(n^{-1})$.
- The single bootstrap bias correction method succeeded in adjusting the bias in the slope of the BAR model estimator.
- The well-known problem "persistent excursion" bias exists and needs more investigation to be adjusted.
- An analytical bias equation for the slope of the BAR model with an intercept has been derived and it is found that it is consistent with the bias in the slope of the LS estimator.
- These conclusions are consistent with the general theory of the autoregression.

In Progress

- **The Persistent Excursion Bias**

Studying the persistent excursion bias is still running and in progress, the results are promising so far.

- **Grid Confidence Interval**

We examined different types of confidence intervals including the asymptotic CI, Percentile CI, the Bias Corrected, and accelerated CI. Although most of them achieve the required coverage, the confidence intervals are not symmetrical.

References

- Cowan, R. and Staudte, R. G. (1986). The bifurcating autoregression model in cell lineage studies. Biometrics, 42: 769–783.
- Chang, J. and Hall, P. (2015). Double-bootstrap methods that use a single double-bootstrap simulation.Biometrika, 102:203–214.
- Elbayoumi, T. and Terpstra, J. (2016). Weighted L_1-estimates for the first-order bifurcating autoregressive model. Communication in Statistics- Simulation and Computation, 45:2991–3013.
- Engsted, T. and Pedersen, T. Q. (2014). Bias-correction in vector autoregressive models: A simulation study. Econometrics, 2(1): 45–71.
- Hall, P. and Horowitz, J. (1996). Bootstrap critical values for tests based on generalized-method-of-moments estimators. Econometrica, 64(4):891–916.
- Kendall, M. G. (1954). A note on the bias in the estimation of an autocorrelation. Biometrikas, 41: 403 – 404.
- Marriott H. C. Pope J. A. (1954). Bias in the estimation of Autocorrelations, Biometrika, Volume 41, Issue 3-4, 3, Pages 390–402.
- Ouysse, R. (2013). A fast iterated bootstrap procedure for approximating the small-sample bias. Communications in Statistics - Simulation and Computation, 42(7): 1472–1494.
- Zhou, J. and Basawa, I. (2005). Least-squares estimation for bifurcating autoregressive processes. Statistics & Probability Letters, 74(1): 77–88.

Thank you