Bias Correction and Machine Learning in AR(1) Estimation: Bridging Traditional and Modern Techniques

Michael M. Müller[1], Günther Specht[1], Lukas Kleinheinz[2] and Janette Walde[2]

[1]Department of Computer Science, Universität Innsbruck, Austria
[2]Department of Statistics, Universität Innsbruck, Austria

Abstract

Autoregressive models are fundamental in time series analysis, with the AR(1) process being particularly relevant in fields like economics for modeling error terms with serial correlation. However, conventional estimation techniques such as Ordinary Least Squares (OLS) and Maximum Likelihood Estimation (MLE) exhibit bias when estimating the AR(1) parameter, especially with short time series data. This bias can impact the reliability of statistical inference when using these methods to model the error term. This paper investigates various bias-correction methods for AR(1), comparing analytical, simulation-based, and bootstrapping techniques in detail. Each method's effectiveness in mitigating bias is assessed, along with a proposed solution to address the overfitting issue highlighted in recent literature, aiming to improve model accuracy. Furthermore, we advocate for exploring machine learning methodologies as a promising approach to enhance AR(1) process estimation. Our findings suggest that the adaptability and ability of machine learning to handle complex patterns could lead to significant advancements in the precision of AR(1) parameter estimates. This innovative approach not only expands the horizons of time series analysis but also creates new avenues for research in econometrics and related fields.

Keywords

Time Series, Machine Learning, Serial Correlation, AR(1), Econometric Analysis, Autoregressive Processes, Bias Correction, Bootstrapping

1. Introduction

The estimation of autoregressive (AR) coefficients is critical in analyzing time series data, often used in economics. Researchers frequently utilize Newey-West heteroscedasticity and autocorrelation consistent (HAC) standard errors to address issues associated with consecutive errors. However, a more advanced option involves modeling error terms using an AR process, usually an AR(1) process, which presents specific benefits over the traditional Newey-West approach, particularly in terms of efficiency of the parameter estimation.

A less recognized issue arises regarding the bias present in frequently used AR coefficient estimators. This bias becomes notably more pronounced when utilized on short time series containing fewer than 50 data points, particularly as the autocorrelation parameter ρ approaches 1, a common phenomenon in macroeconomic research. In these scenarios, the estimator typically underestimates the autocorrelation. Consequently, when researchers choose to model error terms using AR(1) in the presence of serially correlated error terms, there is a significant risk of underestimating autocorrelation. Finally, this results in an increased risk of committing a Type-I error.

This paper addresses the issue of unbiased estimation for AR(1) processes, with a primary emphasis on minimizing bias in parameter estimation while maintaining a reasonable variance.

To address this issue, the first part of this paper offers a comprehensive comparative analysis of bias-reduced AR(1) estimation techniques using an innovative machine learning approach to assess variable importance from random forest models. This contribution fills a notable gap in the existing literature by providing a clear and accessible comparison of these methods.

The main objective of this chapter is to equip researchers with the necessary resources to make well-informed decisions when selecting an appropriate estimation method. The research evaluates three primary approaches commonly used to estimate unbiased AR(1) parameters: simulation, analytical methods, and bootstrapping.

Furthermore, this paper proposes a solution to mitigate overfitting by modifying the simulation-based approach previously introduced by Sørbye et al.[1]

2. Related Work

While there is a significant body of literature on new bias correction techniques for autoregressive (AR) processes [2, 3, 4], a notable gap exists in the evaluation and comparison of the effectiveness of these methods. Some studies have focused on the comparative analysis of bias and empirical standard error estimation in

35th GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken), May 22-24, 2024, Herdecke, Germany.
michael.m.mueller@uibk.ac.at (M. M. Müller)
0009-0005-0918-8833 (M. M. Müller); 0000-0003-0978-7201 (G. Specht); 0000-0002-1241-5292 (J. Walde)
© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-3710/paper4.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

AR(1) models, primarily examining three estimation techniques: Maximum Likelihood Estimation (MLE), Ordinary Least Squares (OLS), and Bayesian methods. Previous research [5, 6] indicated that MLE and Bayesian techniques perform similarly well. However, these studies did not consider the impact of bias correction on these estimators, highlighting a critical gap in our understanding of how bias-reduced estimation techniques compare to traditional approaches.

There also is investigation into strategies for reducing bias in autoregressive model estimation, exploring three techniques: first-order bias correction, bootstrapping, and recursive mean reduction [7]. The paper revealed that bootstrapping was particularly effective in reducing bias, while recursive mean adjustment excelled in reducing mean squared error.

In a related paper the effectiveness of analytical bias correction and bootstrapping in the context of Vector Autoregressive (VAR) models is examined [8]. Their findings showed that for stationary processes, analytical bias correction outperformed bootstrapping, whereas for non-stationary processes, bootstrapping exhibited superior performance over analytical correction. Engsted's research underscores the importance of tailoring bias correction techniques to the specific characteristics of the analyzed data, emphasizing the need for context-aware bias correction methods.

3. Theoretical Background

This section begins with a brief introduction to Autoregressive processes and then explores various methodologies for accurately estimating unbiased autoregressive coefficients of order 1. It offers detailed explanations of the three main approaches commonly used in the field: analytical techniques, simulation-based methods, and bootstrapping. Additionally, it presents a solution to enhance the simulation-based approach proposed by Sørbye et al.

3.1. Autoregressive Processes (AR)

Autoregressive processes are often used as stochastic models to model temporal correlations in economic time series data. These models assume that the current value of a variable is a linear combination of its past values, plus a random error term. AR processes can be represented as AR(p), where p indicates the order of the process, or the number of lagged terms that are included in the model.

Autoregressive (AR) models, especially those of first-order, are common in economics because of their simplicity and effectiveness in capturing persistent patterns. With an order of 1, these models primarily model the current period based solely on the preceding period. Such an AR(1) process can be expressed as follows:

$$y_t = \alpha + \rho y_{t-1} + \epsilon_t \tag{1}$$

where y_t represents the variable of interest at time t, α denotes the intercept term, ρ represents the autoregressive coefficient, y_{t-1} is the lagged value of the variable, and ϵ_t represents the error term.

While AR(1) models provide valuable insights into the dynamics of economic variables, the estimated parameter $\hat{\rho}$ can be subject to bias in certain circumstances when calculated through Yule-Walker, OLS, or Maximum Likelihood methods. This bias is especially present for short time series with less than 50 periods and remains noticeable for up to 100 periods.

Numerous methods exist to achieve less biased results, but the main challenge is navigating the bias-variance trade-off, which the mean squared error (MSE) quantifies as a metric.

$$MSE(\hat{\rho}) = Bias(\widehat{\rho}, \rho)^2 + Variance(\hat{\rho}) \tag{2}$$

When evaluating estimators of the same size based on mean squared error (MSE), it is commonly preferred to opt for an estimator with reduced bias. Nevertheless, it is crucial to account for variance since lower bias does not invariably ensure superior performance. This section explores three primary strategies for attaining unbiased outcomes: analytical methods, simulation-based techniques, and bootstrapping procedures. Furthermore, an enhancement to the most recent simulation-based approach by [1] will be discussed.

3.2. Analytical Approaches

Analytical methods for correcting bias in short-order AR(1) processes have been proposed and studied in the literature. Notable approaches in this area include the work of [9], and the contributions of [10] and [11]. Additionally, [12] demonstrated that the bias of least squares estimators for models of known, finite order is a linear function of the unknown model coefficients, up to order $1/T$.

The analytical approaches aim to develop an expression for the bias in the autoregressive coefficient parameter estimate ($\widehat{\rho}$). The unbiased estimate ($\widehat{\rho_{\text{biascorr}}}$) is computed by deducting the estimated bias from $\widehat{\rho}$. Roy Fuller's approach focuses on addressing unit roots close to 1, making it appropriate for both short-order AR(1) processes and higher-order AR processes.

[11] implements an exact median-unbiased estimator for AR(1) processes, while [10] offer an approximate median-unbiased estimator for higher-order AR processes. Although the latter applies to higher order processes, it may face computational challenges when dealing with high AR orders.

3.3. Bootstrapping Approaches

Bootstrapping approaches have gained popularity as an alternative method for obtaining unbiased estimates in autoregressive processes. Introduced by [13], bootstrapping involves resampling the original dataset with replacement to generate multiple bootstrap samples. These samples allow for the empirical distribution of the estimator to be characterized, providing insights into its variability and bias.

The application of bootstrapping to autoregressive processes was pioneered by [14], who demonstrated the potential of bootstrapping for bias correction in AR models. This method estimates the autocorrelation coefficient ρ from the observed time series and uses this estimate to generate bootstrap replications, effectively simulating the "true" autoregressive process.

Further developments in bootstrapping approaches, such as those proposed by [15] and [16], have focused on improving the accuracy of bias-corrected estimates through innovative techniques like backward AR modeling and residual-based bootstrap methods. These advancements underscore the versatility and effectiveness of bootstrapping in addressing the challenges of bias correction in autoregressive modeling.

3.4. Simulation-Based Approaches

Simulation-based methods represent a powerful tool for correcting bias in autoregressive coefficient estimation, especially for AR(1) processes. These approaches, as detailed by the recent research of Sørbye et. al [1], employ computational simulations to model the true parameter ρ as a function of the originally biased estimate $\widehat{\rho}$. The essence of these methods lies in their ability to utilize a vast array of simulated time series, where the true parameter ρ is known, allowing for the direct modeling and understanding of bias in the estimated coefficients.

A notable technique within simulation-based approaches is the use of Hermite polynomials of order 3 to model the relationship between the biased estimate $\widehat{\rho}$ and the true parameter ρ. This method involves normalizing the ρ coefficients within the stationary interval $(-1, 1)$ to $[0, 1]$ and applying a logit transformation to ensure that the corrected coefficients remain within the stationary range:

$$g(\rho) = \text{logit}\left(\frac{\rho + 1}{2}\right) \tag{3}$$

The corrected estimate $\widehat{\rho_{\text{biascorr}}}$ is then obtained through a function f, parameterized by a vector $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$, representing the coefficients of the Hermite polynomials:

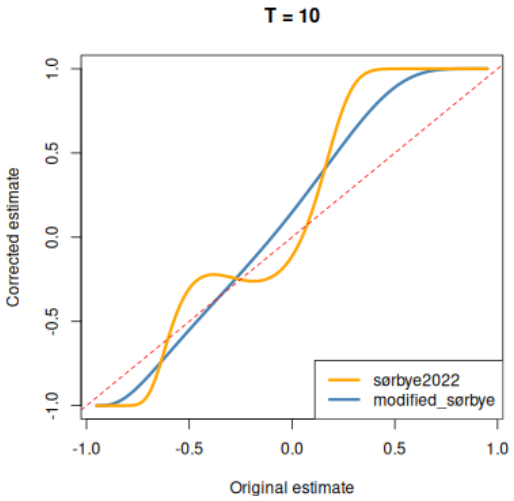

Figure 1: Correction Curve for a time series with 10 time periods showing severe overfitting for the Sørbye approach. Solved by our modified version.

$$\widehat{\rho_{\text{biascorr}}} = f(\widehat{\rho}) = \beta_0 + \beta_1\widehat{\rho} + \beta_2(\widehat{\rho}^2 - 1) + \beta_3(\widehat{\rho}^3 - 3\widehat{\rho}) \tag{4}$$

The optimization of β values is achieved through minimizing the weighted squared error between the bias-corrected estimate and the true parameter across a finely spaced grid of ρ values. This process ensures that the corrected estimates are as close as possible to the true parameter values, thereby reducing bias.

4. Modified Sørbye Approach

Despite the effectiveness of the Sørbye et al. approach, challenges such as overfitting become apparent when applied to short time series (10-15 periods). This overfitting is particularly noticeable in the correction curves, where the adjusted values deviate significantly from the expected outcomes, suggesting a misalignment in the bias correction process. To address this issue, we propose a modification to the Sørbye et al. approach. Instead of calculating the mean, we propose recalculating the median during the optimization process of the Hermite polynomial parameters:

$$\widehat{\beta_T} = \underset{\beta}{\text{argmin}} \sum_{r=1}^{l} \left|\text{median}_{j=1}^{m}\left(f(\widehat{\rho}_{rj}, \beta) - \rho_r\right)\right| \tag{5}$$

This adjustment aims to mitigate overfitting by leveraging the robustness of the median to outliers, thereby

providing a more accurate correction curve for short time series. The modified approach extends the applicability of the simulation-based bias correction to time series ranging from 5 to 100 periods, offering a comprehensive solution for bias correction across various time series lengths. We present an adapted version of the Sørbye et al. approach available through the R package provided at https://github.com/michael-mueller-uibk/ar1MedianBiascorrection.

5. Analysis and Results

Simulations for the comparative analysis were conducted following a systematic procedure. To simulate the process, a precise grid of parameter values for ρ were used, ranging from -0.99 to 0.99 in increments of 0.01. In addition, these simulations were performed across various time periods denoted as T. To ensure reliable results, 1000 simulated time series were undertaken for each specified ρ value. Table 1 shows the estimation techniques that were evaluated.

This process led to the development of a varied dataset that comprises time series data featuring the actual parameter ρ and its respective estimated values derived from earlier mentioned estimators. We used this dataset to carry out a comprehensive empirical analysis, which facilitated the computation of critical statistical metrics like empirical bias, variance, and mean squared error. We used these metrics as important indicators to evaluate the performance of the estimators being considered.

Estimation Technique	Type
Kim [15]	Bootstrapping
Roy-Fuller [9]	Analytical
Shaman-Stine [12]	Analytical
Modified Sørbye	Simulation based

Table 1
Compared Estimation Techniques.

An alternative method for comparing various bias correction techniques is presented in this section. The distinct capabilities of bias correction when applied to short and long time series and their unequal performance with positive and negative ρ values have led to the development of a Random Forest ensemble estimator for ρ. The objective is to capitalize on the advantages of individual estimators and attain superior overall outcomes by generating a collection of decision trees that are trained on a set of simulated time series where the true parameter ρ is known and the corresponding estimators $\hat{\rho}$. The random forest imposes arbitrary limitations on each tree, creating a variance reduction in the forest when the forest estimator is computed through the average, given that each forecast differs.

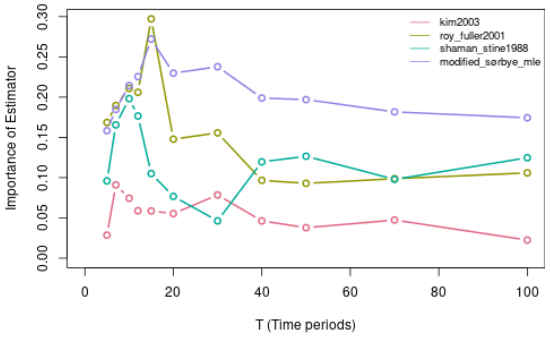

Figure 2: Comparison of Estimator Performance. The figure illustrates the importance of estimators, showing the consistently high performance of our modified Sørbye approach compared to the analytical and bootstrapping methods.

The Random Forest can be utilized to assess the overall performance of the estimators. While Random Forests are often perceived as complex 'black box' models because of their intricacy, a method exists for quantifying the significance of the variables they employ. To do so, a Random Forest model is trained for each time period T using the previously generated simulations. In this process, the true autocorrelation parameter ρ is used as the dependent variable, with the explanatory variables comprising the estimators acquired from the different biascorrection techniques. Variable importance is evaluated by measuring how much an estimator contributes to the predictive accuracy of the model. This assessment is conducted by observing the change in the model's prediction error when the values of a specific attribute are randomly shuffled. If shuffling an estimator's values leads to a discernible increase in the model's prediction error, it indicates that the model heavily depends on that estimator for its predictions, categorizing the attribute as 'important'. Conversely, if reordering the estimator's values has little impact on the model's error, it suggests that the model does not heavily depend on that estimator, rendering it 'unimportant' in the prediction process. The concept of permutation estimator importance, originally introduced in the context of Random Forests by [17], is applied in this paper.

To evaluate the estimators' performance, the permutation variable importance for each Random Forest is computed. Figure 2 displays these results. It is evident that the modified Sørbye approach consistently exhibits the highest variable importance across all time periods (T), except for when $T = 15$. Moreover, the two analytical approaches demonstrate strong performance. Notably, the Roy-Fuller [9] estimator displays higher importance for shorter time series, while the Shaman-Stine [12] estimator performs slightly better for longer time series. In

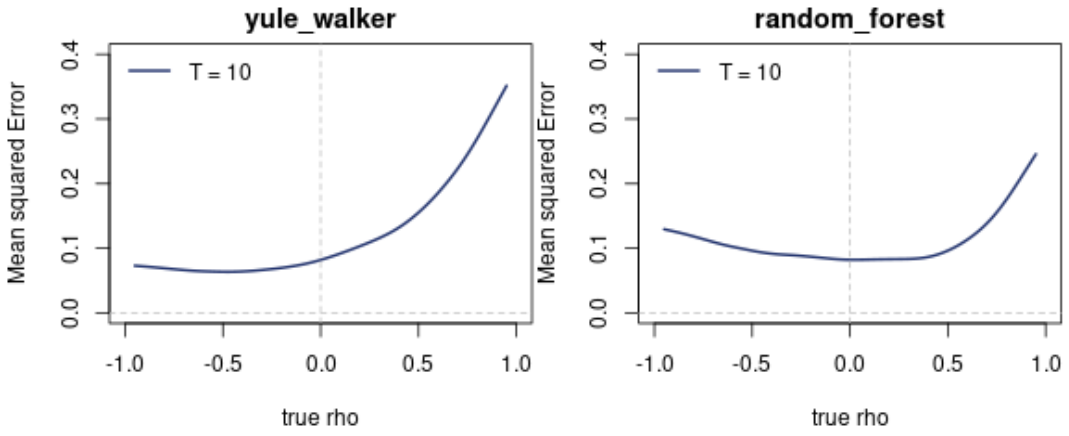

Figure 3: Significant enhancement in mean squared error seen in a 10-period time series by tuning the Yule-Walker parameter with a random forest

contrast, the bootstrapping approach by Kim [15] consistently shows the lowest variable importance out of all the estimators.

The variable importance plot results should not be considered conclusive evidence of the effectiveness of biascorrection methods. Caution must be used in interpreting the data because if two estimators are extremely similar, they may have overlapping information, resulting in a relatively unchanged prediction error when one is permuted. Conversely, if the estimators contain different information, it may result in a more significant increase in prediction error.

[18] provides a more comprehensive view of the significance of the estimators by acknowledging that certain estimation methods may equally fit the data. Their approach and the transfer to our problem of comparing the estimation approaches for our data is out of scope of this paper, but should be considered for future comparisons of the approaches.

6. Future Work

This section examines the application of a random forest model for bias correction. The goal is to decrease the bias of the simple to compute Yule-Walker estimate, therefore we simulate time series of length T with known true parameter ρ. We then compute the Yule-Walker estimate for each of these time series. The Yule-Walker estimate and the length of the time series T can then be utilized as explanatory variables in the random forest, along with the true parameter ρ as the dependent variable. Although this approach provides initial insights into the potential of incorporating machine learning techniques into the bias correction of AR processes, further enhancements in performance and reductions in bias and mean squared error are possible through the use of a more comprehensive ensemble estimator that integrates maximum likelihood estimation and bias-corrected estimates from different models.

Figure 3 shows the outcomes, demonstrating the impact of the random forest on the mean squared error. The results highlight that the random forest can effectively decrease bias of the Yule-Walker estimate for larger ρ values, although there is a minor rise in bias for negative ρ values. Additionally, it shows the technique's proficiency in reducing variance for positive ρ values. These findings suggest opportunities for enhancing the estimators' performance through further investigation of advanced ensemble techniques incorporating more bias-corrected estimators.

7. Conclusion

This paper contributes to the field by addressing the challenge of estimating AR(1) coefficients in short time series. The paper explores various strategies for mitigating bias in AR(1) parameter estimation and introduces an effective adaptation of the bias correction methodology initially proposed by Sørbye et al. [1]. The research findings suggest that analytical and simulation-based methods are more effective for estimating stationary AR(1) processes compared to bootstrapping, aligning with conclusions from prior studies [8].

The newly proposed modified Sørbye approach demon-

strates promise in estimating AR(1) parameters by reducing bias while maintaining low variance. Additionally, analytical techniques proposed by Roy et al. [9] and Shaman et al. [12] are identified as viable alternatives for researchers, showing comparable performance and versatility for higher-order AR processes.

In conclusion, this paper emphasizes the significance of bias correction in AR coefficient estimation and highlights the importance of tailored bias correction methods that consider the specific characteristics of the analyzed data.

Moreover, the paper suggests the potential benefits of employing machine learning approaches to enhance AR estimation, opening avenues for further research and methodological advancements.

References

[1] S. H. Sørbye, P. G. Nicolau, H. Rue, Finite-sample properties of estimators for first and second order autoregressive processes, Statistical Inference for Stochastic Processes 25 (2022) 577–598.

[2] T. Elbayoumi, S. Mostafa, Impact of bias correction of the least squares estimation on bootstrap confidence intervals for bifurcating autoregressive models, Journal of Data Science (2023) 25–44. doi:10.6339/23-JDS1092.

[3] J. Breitung, S. Kripfganz, K. Hayakawa, Bias-corrected method of moments estimators for dynamic panel data models, Econometrics and Statistics 24 (2022) 116–132. URL: https://www.sciencedirect.com/science/article/pii/S2452306221000770. doi:https://doi.org/10.1016/j.ecosta.2021.07.001.

[4] W. Na, C. Yoo, Real-time bias correction of beaslesan dual-pol radar rain rate using the dual kalman filter, Journal of Korea Water Resources Association 53 (2020) 201–214.

[5] T. Krone, C. J. Albers, M. E. Timmerman, Comparison of estimation procedures for multilevel ar(1) models, Frontiers in Psychology 7 (2016). doi:10.3389/fpsyg.2016.00486.

[6] T. Krone, C. J. Albers, M. E. Timmerman, A comparative simulation study of AR(1) estimators in short time series, Quality & Quantity 51 (2017) 1–21. doi:10.1007/s11135-015-0290-1.

[7] K. D. Patterson, Bias reduction through first-order mean correction, bootstrapping, and recursive mean adjustment, Journal of Applied Statistics 34 (2007) 23–45. doi:10.1080/02664760600994638.

[8] T. Engsted, T. Q. Pedersen, Bias-correction in vector autoregressive models: A simulation study, Econometrics 2 (2014) 45–71. doi:10.3390/econometrics2010045.

[9] A. Roy, W. A. Fuller, Estimation for autoregressive time series with a root near one, Journal of Business & Economic Statistics 19 (2001) 482–493.

[10] D. W. K. Andrews, H.-Y. Chen, Approximately Median-Unbiased Estimation of Autoregressive Models, Journal of Business & Economic Statistics 12 (1994) 187–204. doi:10.1080/07350015.1994.10510007.

[11] D. W. K. Andrews, Exactly median-unbiased estimation of first order autoregressive/unit root models, Econometrica 61 (1993) 139. doi:10.2307/2951781.

[12] P. Shaman, R. A. Stine, The bias of autoregressive coefficient estimators, Journal of the American Statistical Association 83 (1988) 842–848.

[13] B. Efron, Computers and the theory of statistics: Thinking the unthinkable, SIAM Review 21 (1979) 460–480. doi:10.1137/1021092. arXiv:https://doi.org/10.1137/1021092.

[14] R. A. Stine, P. Shaman, A fixed point characterization for bias of autoregressive estimators, The Annals of Statistics 17 (1989) 1275–1284.

[15] J. H. Kim, Forecasting autoregressive time series with bias-corrected parameter estimators, International Journal of Forecasting 19 (2003) 493–502.

[16] H. Tanizaki, S. Hamori, Y. Matsubayashi, On least-squares bias in the AR(p) models: Bias correction using the bootstrap methods, Statistical Papers 47 (2006) 109–124. doi:10.1007/s00362-005-0275-6.

[17] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.

[18] A. Fisher, C. Rudin, F. Dominici, All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously, J. Mach. Learn. Res. 20 (2019) 1–81.