Testing Autocorrelation and Partial Autocorrelation: Asymptotic Methods versus Resampling Techniques

Zijun Ke
Sun Yat-Sen University

Zhiyong (Johnny) Zhang
University of Notre Dame

Ke, Z., & Zhang, Z. (2018). Testing Autocorrelation and Partial Autocorrelation: Asymptotic Methods versus Resampling Techniques. British Journal of Mathematical and Statistical Psychology, 71(1), 96–116.

This research was supported by the grant program on Statistical and Research Methodology in Education from the Institute of Education Sciences of U.S. Department of Education (Grant no: R305D140037).

Author Note

Correspondence should be addressed to Zijun Ke, Department of Psychology, Sun Yat-Sen University, Higher Education Mega Center, Guangzhou, Guangdong 510006, China. Email: keziyun@mail.sysu.edu.cn.

Abstract

Autocorrelation and partial autocorrelation, which provide a mathematical tool to understand repeating patterns in time series data, are often used to facilitate the identification of model orders of time series models, e.g., moving average (MA) and autoregressive (AR) models. Asymptotic methods for testing autocorrelation and partial autocorrelation such as the $1/T$ approximation method and the Bartlett's formula method may fail in finite samples and are vulnerable to nonnormality. Resampling techniques such as the moving block bootstrap and the surrogate data method are competitive alternatives. In this study, we used a Monte Carlo simulation study and a real data example to compare asymptotic methods with the aforementioned resampling techniques. For each resampling technique, we considered both the percentile method and the bias-corrected and accelerated method for interval construction. Simulation results showed that the surrogate data method with percentile intervals yielded better performance than the other methods. An R package `pautocorr` is developed and demonstrated to carry out tests evaluated in this study.

Testing Autocorrelation and Partial Autocorrelation: Asymptotic Methods versus Resampling Techniques

Growing interest in studying behavioral change via time series analysis has been observed in social and behavioral sciences (Browne & Nesselroade, 2005; Borckardt et al., 2008; Ferrer & Zhang, 2009). By collecting repeated observations of the same individual, researchers are able to study patterns of change unique to the individual and model lagged effects of the same variable or among variables. Among the classic time series models for delineating lagged linear relations are autoregressive models (AR), moving average (MA) models, and the combination of these two models, i.e., autoregressive moving average (ARMA) models (e.g., Shumway & Stoffer, 2006). In an AR model of order p, abbreviated as AR(p), the current value of the series can be predicted as a function of the p past values. Likewise, in an MA model of order q, abbreviated as MA(q), the current value of the series can be predicted as a function of the q past shock variables (unpredicted factors). As the integration of AR and MA models, the current value of an ARMA model can be explained as a function of the p past values and q past shock variables.

To build ARMA models, one key step is to identify model orders, i.e., to get preliminary values of p and q. Traditionally, researchers determine the values of p and q by testing the significance of the autocorrelation / partial autocorrelation function (ACF / PACF). Similar to correlation coefficients for independent data, the ACF is the correlation between the current value of the series and values at different time points. In other words, it quantifies the similarity between observations as a function of the time lag between them. The PACF, as implied by its name, is a "partial correlation" version of the ACF. Specifically, the PACF is the correlation between the current value of the series and values at different time points after partialling out the effects of the values in-between. Statisticians have found that the behavior of the ACF and PACF of stationary and invertible ARMA models[1] is linked to the values of p and q. More specifically,

[1] There are strict and weak forms of stationarity. Strict stationarity ensures that the cumulative distribution function of every collection of values is invariant across time points. Weak stationarity means that statistical properties such as mean, variance, and covariance agree with their counterparts in the shifted sets. In this paper, stationarity refers to the latter. Invertibility ensures model uniqueness. Two different MA (ARMA) models can produce the same time series (e.g., Shumway & Stoffer, 2006, pp. 91-92; du Toit & Browne, 2007, Tables 6-7). In particular, one of these MA (ARMA) models can be "inverted" to be an AR model of an infinite order and thus is defined as being invertible. In this light, invertibility helps solving the problem of nonuniqueness. More formally, an ARMA(p, q) model is stationary

for AR(p) models, the ACF tails off and the PACF cuts off after lag p; for MA(q) models, the PACF tails off and the ACF cuts off after lag q; and for ARMA(p, q) models, both the ACF and PACF tail off. Making use of the association between model orders and the (P)ACF, researchers can determine the values of p and q by studying the behavior of the (P)ACF. However, in real data analysis, the ACF and PACF are sample estimates. Consequently, researchers have to rely on hypothesis testing to examine the significance of the ACF and PACF.

Two well-established asymptotic methods for testing autocorrelations are the $1/T$ approximation (e.g., Shumway & Stoffer, 2006, p.519-520) and the Bartlett's formula methods (Bartlett, 1955, p.289; Shumway & Stoffer, 2006, p.519-520; Zhang & Browne, 2010). The $1/T$ approximation method has been adapted for partial autocorrelations whereas the Bartlett's method has not. Given the limitations of asymptotic methods in small samples (e.g., unwarranted finite sample performance and the vulnerability to nonnormality in small samples), resampling techniques have been developed for assessing (partial) autocorrelations. Usually, the bootstrap comes into play when an analytical solution is difficult to be adapted to small samples. To address the problem of the lack of independence in time series data, the moving block bootstrap has been proposed for inference on (partial) autocorrelations (Künsch, 1989; Efron & Tibshirani, 1994, p.99-102). While the bootstrap resamples from a distribution assumed to generate the observed data, the surrogate data method resamples under the null hypothesis (Theiler et al., 1992). In other words, the bootstrap method tries to preserve the time dependence of the data whereas the surrogate data method destroys the time dependence to resemble the null. The jackknife is another commonly used resampling technique for estimating the bias, the standard error and the influence function (Efron, 1994). To study the bias and the variability of the estimator, the jackknife creates replications by sequentially deleting one or more observations. Because only one or a few observations are altered, estimates based on the resulting replications are correlated with each other and thus the variability (difference) across jackknife replications is much smaller

and invertible only when the roots of $A(z) = 1 - \sum_{j=1}^{p} a_j z^j$ and $B(z) = 1 + \sum_{j=1}^{q} b_j z^j$ with a_js and $b_j s$ being AR and MA weights respectively lie outside the unit circle. Practically, with population a_js and b_js, one may check the stationarity and invertibility of an ARMA model by constructing the $\boldsymbol{A}$ matrix in du Toit & Browne (2007, Equ. 16) using a_js or b_js and examining whether the absolute eigenvalues of the matrix are less then one. In finite samples, the augmented Dickey-Fuller test is often used to test whether differencing is needed to make the observed series stationary (e.g., Brockwell & Davis, 2002, pp. 193-196).

than that of the sampling distribution. To take into account the non-independence, the jackknife estimators of the bias and the standard error use a correction factor to enlarge the difference or the variability. Because for different statistics, the correction factor may or may not be valid, the bootstrap is usually more general than the jackknife.

Although how to determine model orders of ARMA models has been extensively discussed in the literature, two issues have complicated the implementation of these existing methods in reality. First, small sample size and nonnormality pose a threat to the validity of asymptotic methods. Sample size in psychological studies is usually small to moderate. Thus asymptotic methods that are based on large sample behavior may fail in small samples. In addition, asymptotic methods usually assume normality to achieve mathematical simplicity. However, psychological data rarely follow normal distributions (Micceri, 1989). Therefore, whether asymptotic methods for testing (partial) autocorrelations are robust to small sample size and nonnormality awaits further investigation.

Second, although bootstrapping (partial) autocorrelations to make inference has been largely relied on the moving block bootstrap, it is not necessary that the moving block bootstrap is superior to the surrogate data method when the interest is to test (partial) autocorrelations. A related study on tests of correlations for independent data showed that univariate bootstrap which resamples under the null outperformed bivariate bootstrap which resamples under the alternative when testing correlations (Lee & Rodgers, 1998). In addition, among the several interval construction methods for resampling techniques, bias-corrected and accelerated (BCa) intervals are often recommended because they are second-order accurate, as opposed to the first order accurate percentile intervals (Efron & Tibshirani, 1994, p.187)[2]. As we will show in this study, BCa intervals may not always be superior to percentile intervals. This exception is likely related to the computation of the bias-correction factor during the construction of BCa intervals for the surrogate data method.

In sum, because the relative performance of the aforementioned tests of (partial) autocorrelations is largely unknown, the objective of this study is 1) to compare the relative finite sample performance of various tests (i.e., the $1/T$ approximation method, the Bartlett's formula

[2] Second-order accurate means that the error in matching the intended miss-coverage rates goes to zero at a rate of $1/T$ in terms of sample size or time series length T. For first-order accurate, the rate is $1/\sqrt{T}$.

method, the moving block bootstrap and the surrogate data method) of (partial) autocorrelations with normal and nonnormal data, 2) to develop an easy-to-use R package `pautocorr` to implement the aforementioned tests, and 3) to provide a practical guideline on applications of those tests under various circumstances.

The rest of the article is organized as follows: the two asymptotic tests are described first, followed by a section on resampling techniques. A simulation study is presented to evaluate the performance of those tests. An R package and a real data example are included to illustrate the use of different tests of (partial) autocorrelations. Finally, this article ends with concluding comments.

Asymptotic Methods

The $1/T$ approximation method for (partial) autocorrelation testing is motivated by the fact that under the null hypothesis which assumes that all autocorrelations are zero, time series data reduce to uncorrelated data. Therefore, the well-established standard error estimator for correlation, $1/\sqrt{T}$ where T is the sample size, can be used directly for (partial) autocorrelations.

More formally, let y_1, y_2, . . ., y_T be a stationary time series of length T whose statistical properties such as mean, variance and covariance are invariant across different time points. The concurrent variance $\gamma_0 = \text{Var}\left(y_t\right)$ and the lagged covariances $\gamma_l = \text{Cov}\left(y_{t+l}, y_t\right) = \text{E}\left[\left(y_{t+l} - \mu\right)\left(y_t - \mu\right)\right]$ are estimated by their sample counterparts using

$$\hat{\gamma}_l = \frac{1}{T}\sum_{t=1}^{t=T-l}\left(y_{t+l} - \bar{y}\right)\left(y_t - \bar{y}\right)'$$

where $\bar{y}$ is the sample mean $\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$. The $1/T$ approximation method assumes that under the null, autocorrelation estimates follow a normal distribution $\hat{\rho}_l \sim N\left(0, \frac{1}{T}\right)$ where $\rho_l = \frac{\gamma_l}{\gamma_0}$ is the population autocorrelation at lag l and $\hat{\rho}_l = \frac{\hat{\gamma}_l}{\hat{\gamma}_0}$ is its sample counterpart.

The partial autocorrelation ϕ_l for stationary series is defined as $\phi_1 = \rho_1$ and $\phi_l = \text{Corr}\left(y_t - y_t^{t-l+1}, y_{t-l} - y_{t-l}^{t-l+1}\right)$ for $l \geq 2$. Here y_t^{t-l+1} denotes the best linear prediction of y_t based on $y_{t-1}, y_{t-2}, \ldots, y_{t-(l-1)}$[3]. In plain language, ϕ_l is the correlation between y_t and y_{t-l} after removing the association attributable to $y_{t-1}, y_{t-2}, \ldots, y_{t-(l-1)}$. Under the null, ϕ_l reduce to ρ_l because both y_t and y_{t-l} are unrelated with $\left\{y_{t-1}, y_{t-2}, \ldots, y_{t-(l-1)}\right\}$:

[3] Theoretically, y_t^{t-l+1} can be obtained by regressing y_t on $y_{t-1}, y_{t-2}, \ldots, y_{t-(l-1)}$ in the population and calculating the predicted value for y_t using the obtained regression coefficients.

$\phi_l = \text{Corr}\left(y_t - y_t^{t-l+1}, y_{t-l} - y_{t-l}^{t-l+1}\right) = \text{Corr}\left(y_t, y_{t-l}\right) = \rho_l$. In other words, under the null that data points are uncorrelated with each other, the sampling distribution of $\hat{\phi}_l$, the estimate of ϕ_l, should be the same as that of $\hat{\rho}_l$. Hence, the $1/T$ approximation method assumes the same distribution for $\hat{\phi}_l$, $\hat{\phi}_l \sim N\left(0, \frac{1}{T}\right)$. Put it together, the $1/T$ approximation method computes the test statistic $\frac{\hat{\rho}_l}{1/\sqrt{T}}$ or $\frac{\hat{\phi}_l}{1/\sqrt{T}}$ and compares it to a critical value of $Z_{\alpha/2}$ or $Z_{1-\alpha/2}$ where $Z \sim N(0,1)$ and α is the chosen significance level.

Because partial autocorrelations remove the dependence between y_t and y_{t-l} due to $\left\{y_{t-1}, y_{t-2}, \ldots, y_{t-(l-1)}\right\}$, estimating ϕ_l is not as straightforward as estimating ρ_l. The Durbin-Levinson algorithm is a widely used iterative algorithm for computing $\hat{\phi}_l$ for stationary series. Let ϕ_{lk} denotes the regression coefficient of y_{t-k} in the one-step-ahead prediction $y_t \sim \phi_{l1} y_{t-1} + \phi_{l2} y_{t-2} + \ldots + \phi_{ll} y_{t-l}$. By the property of best linear prediction, stationarity, and matrix algebra, it can be shown that $\phi_{ll} = \phi_l$[4]. The Durbin-Levinson algorithm is used to compute the regression coefficients in the one-step-ahead prediction. In particular, with $\hat{\phi}_0 = 0$, for $l \geq 1$, the algorithm computes $\hat{\phi}_l$ as well as other regression coefficients as follows (e.g., see Shumway & Stoffer, 2006, pp.113-114),

$$\hat{\phi}_{ll} = \hat{\phi}_l = \frac{\hat{\rho}_l - \sum_{k=1}^{l-1} \hat{\phi}_{l-1,k} \hat{\rho}_{l-k}}{1 - \sum_{k=1}^{l-1} \hat{\phi}_{l-1,k} \hat{\rho}_k}$$

where for $l \geq 2$ and $k = 1, 2, \ldots l-1$,

$$\hat{\phi}_{lk} = \hat{\phi}_{l-1,k} - \hat{\phi}_l \hat{\phi}_{l-1,l-k}.$$

Clearly, the Durbin-Levinson algorithm computes $\hat{\phi}_l$ using sample autocorrelations $\hat{\rho}_l$ rather than raw data. In this study, when raw series are not available, we rely on the Durbin-Levinson algorithm to obtain sample partial autocorrelations[5].

The Bartlett's formula was originally developed for obtaining the asymptotic distribution for autocorrelations under general conditions, not only under the null hypothesis. According to the Bartlett's formula (Bartlett, 1955, p. 289), the asymptotic distribution of $\sqrt{T}\left(\hat{\rho}_l - \rho_l\right)$ is $N\left(0, W_l\right)$ where

$$W_l = \sum_{u=-\infty}^{\infty} \left\{2\rho_l^2 \rho_u^2 - 2\rho_l \rho_u \left(\rho_{u+l} + \rho_{u-l}\right) + \rho_u^2 + \rho_{u-l}\rho_{u+l}\right\}. \tag{1}$$

[4] Detailed discussion can be found in Exercise 3.12 in Shumway & Stoffer (2006).

[5] The first author has tested that the resulting sample partial autocorrelations were exactly the same as those obtained from the built-in function `pacf` in the widely used statistical software R.

In short, tests based on the Bartlett's formula first construct the confidence interval for ρ_l using $\hat{\rho}_l \pm Z_{\alpha/2}\sqrt{\frac{\hat{W}_l}{T}}$ where $\hat{W}_l$ is obtained by replacing population autocorrelations in W_l with their sample counterparts, and then examine whether the test value, 0, falls into the obtained confidence interval. One practical concern of this method is that in Equation (1) the asymptotic variance of $\hat{\rho}_l$ is a function of concurrent and lagged correlations up to infinity. In practice, we need to set an upper limit for the number of summed terms. Studies indicated that 30 works satisfactorily (Zhang & Browne, 2010). Thus in this study, we use 30. According to our knowledge, the Bartlett's formula has not been adapted for partial autocorrelations. Thus, when evaluating tests of partial autocorrelations, we would not consider the Bartlett's formula method.

Resampling Techniques

Requiring fewer assumptions and spurred by the advance of computer technology, resampling techniques have become increasingly popular. Two resampling techniques proposed for testing autocorrelation and partial autocorrelation are the bootstrap method and the surrogate data method. Because these two methods often make statistical inference through confidence intervals, we first briefly review two widely used interval construction methods. We then move on to the bootstrap method. The surrogate data method is next.

Interval Construction Methods

Two widely used interval construction methods for resampling techniques are the percentile method and the bias-corrected and accelerated (BCa) method. Interested readers can refer to Efron & Tibshirani (1994) and Efron (1987) for detailed discussion. Here, we summarize the main ideas from the book and the paper. Let $\hat{\theta}^{\star(\alpha/2)}$ be the $100\frac{\alpha}{2}$th percentile of (partial) autocorrelation estimates from B resampling replications. The percentile interval of intended coverage $1-\alpha$ is obtained by finding the corresponding percentiles, i.e., $[\hat{\theta}^{\star(\alpha/2)}, \hat{\theta}^{\star(1-\alpha/2)}]$. For example, suppose that we decide to use a certain resampling procedure, e.g., the bootstrap. Following this procedure, we form a replication sample and we compute the estimate for this replication. This process is repeated for 2000 times and 2000 estimates are obtained. The percentile interval endpoints are the 50th and the 1950th ordered values of the 2000 estimates.

According to Efron (1987), percentile intervals are ideal for statistics in which there exists a monotonic transformation g such that $\frac{g(\hat{\theta})-g(\theta)}{\sigma_{g(\hat{\theta})}} \sim N(0,1)$ where $\sigma_{g(\hat{\theta})}$ is the constant standard error of $g\left(\hat{\theta}\right)$. A related example is the sample correlation coefficient for bivariate normally distributed data. The sampling distribution of a sample correlation is skewed if the true correlation is far away from zero. In addition, the standard error of the sample correlation gets larger as the true correlation gets closer to zero. Fisher's z transformation is one of the g transformation mentioned above because it normalizes the sampling distribution (the transformed statistic follows a normal distribution) and stabilizes the standard error (the standard error becomes independent of the true correlation). Analytically, researchers may find out the interval endpoints for $g\left(\hat{\theta}\right)$ first, and then transform them back to the metric of $\hat{\theta}$ to obtain interval estimates for $\hat{\theta}$. As pointed out by Efron (1987), the bootstrap automatically goes through this process and there is no need to find out the exact transformation g.

BCa intervals, however, are preferable to percentile intervals when the transformation g can only normalize the statistic but fail to correct bias and stabilize the standard error. More specifically, the transformed statistic $g\left(\hat{\theta}\right)$ is allowed to have bias and a non-constant standard error, $\frac{g(\hat{\theta})-g(\theta)}{\sigma_{g(\hat{\theta})}} \sim N\left(-z_0\sigma, \sigma^2\right)$ where $\sigma = 1 + cg(\theta)$ (Efron, 1987). The two quantities z_0 and c are called bias and acceleration constants. Obviously, z_0 is proportional to the difference between the expectation (or median) of $g\left(\hat{\theta}\right)$ and $g(\theta)$. This is how z_0 gets its name. To estimate z_0, Efron & Tibshirani (1994) suggested assessing the amount of median bias by counting the proportion of replications less than the original estimate. The corresponding percentile in the standard normal distribution reflects the size of median bias on the scale of the standard normal distribution and thus can be used as $\hat{z}_0$, the estimate of z_0. Similarly, acceleration c gets its name because it is the rate of change of the standard error with respect to the true parameter value. Efron (1987) proposed to approximate c by one-sixth of the "skewness" of the influence function. The influence function generally measures how the estimator changes if the distribution of data changes slightly. Because the jackknife slightly "changes" the empirical distribution of data by deleting one or a few observations sequentially, there is a close connection between the jackknife and the influence function. Thus, the jackknife is often used to estimate the influence function and hence is recommended for obtaining $\hat{c}$, the estimate of c (Frangos & Schucany, 1990). Let $\boldsymbol{y}_{-t}$ be the original sample with y_t omitted and $\hat{\theta}_{(t)}$ be the estimate of (partial) autocorrelations from sample

$\boldsymbol{y}_{-t}$. Define $\hat{\theta}_{(\cdot)} = \sum_{t=1}^{T} \hat{\theta}_{(t)}/T$. The acceleration $\hat{c}$ can be obtained using the following equation

$$\hat{c} = \frac{\sum_{t=1}^{T} \left(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(t)}\right)^3}{6\left[\sum_{t=1}^{T} \left(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(t)}\right)^2\right]^{3/2}}.$$

With $\hat{z}_0$ and $\hat{c}$, the BCa interval of intended coverage $1-\alpha$ is obtained by finding the following percentiles $[\hat{\theta}^{\star(\alpha_L)}, \hat{\theta}^{\star(\alpha_U)}]$ where

$$\begin{aligned} \alpha_L &= \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha/2)}}{1 - \hat{c}\left(\hat{z}_0 + z^{(\alpha/2)}\right)}\right) \\ \alpha_U &= \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha/2)}}{1 - \hat{c}\left(\hat{z}_0 + z^{(1-\alpha/2)}\right)}\right). \end{aligned}$$

Here $z^{(\alpha/2)}$ is the $100\alpha/2$th percentile of the standard normal distribution and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

When $z_0 = c = 0$, BCa intervals reduce to percentile intervals. One may expect comparable performance of these two methods. When z_0 and c are nonzero, Efron (1987) has shown that BCa intervals are second-order accurate whereas percentile intervals are only first-order accurate. In other words, empirical miss-coverage rates of BCa intervals converge to the intended coverage rate faster than do percentile intervals as the time series length increases. In the example of correlation, if the true correlation is zero, whether or not the bivariate normality assumption is violated, Fisher z transformation of sample correlation approximately shows zero bias and its standard error is independent of the true correlation value (e.g., Hawkins, 1989). Hence, in this situation, percentile intervals are comparable to BCa intervals, provided that our estimates of z_0 and c are accurate. When the true correlation is nonzero and the bivariate normality assumption is violated, the standard error of the transformed correlation depends on the true correlation value (e.g., Hawkins, 1989). In this situation, BCa intervals are likely better than percentile intervals. Similar results can be expected for autocorrelations and partial autocorrelations.

Moving Block Bootstrap

The bootstrap, originally proposed for independent data, is a computer based method of statistical inference that can avoid the formidable wall of mathematics (Efron & Tibshirani, 1994): It samples with replacement from the original data points to form bootstrap samples and makes inference of the original sample based on bootstrap samples in order to model the inference from sample data to a population. However, this naive bootstrap cannot be directly applied to time series data because resampling with replacement can destroy the dependence in time series data.

Because autocorrelations of time series data are often substantial at initial lags but quickly decay to zero, the dependence among observations can be captured by blocks consisting of sufficient consecutive observations. In accordance with this property, moving block bootstrap splits the original time series data into $T - b + 1$ consecutive blocks of length b: y_1 to y_b will be block 1, y_2 to y_{b+1} will be block 2, and etc. (Efron & Tibshirani, 1994, p.99-102). These $T - b + 1$ blocks will then be resampled with replacement. Finally, a bootstrap sample is formed by aligning the selected T/b blocks.

One limitation of this "simple" moving block bootstrap is that points connecting different blocks are bad joints[6]. To avoid this problem, Künsch (1989) proposed the vectorized moving block (VMB) bootstrap. This bootstrap forms blocks in the same way as does the simple moving block bootstrap. The difference occurs when computing (partial) autocorrelations for a bootstrap sample. Unlike the simple moving block bootstrap which treats the bootstrap series as the original data and computes sample (partial) autocorrelations as in the original sample, the VMB bootstrap computes sample autocorrelations based on pairs of observations l lags distant apart within each block, when the parameter of interest is autocorrelation at lag l. After obtaining the first several autocorrelation estimates, sample partial autocorrelations can be obtained via the Durbin-Levinson algorithm. For example, suppose autocorrelation at lag 1 is of interest and the first block consists of observations y_1, y_2, and y_3, and the second block consists of y_8, y_9, and y_{10}. Then the VMB bootstrap computes an estimate of autocorrelation at lag 1 using pairs of observations, $(y_{1,}y_2)$, $(y_{2,}y_3)$, $(y_{3,}y_4)$, $(y_{8,}y_9)$, $(y_{9,}y_{10})$, and $(y_{10,}y_{11})$. And according to the Durbin-Levinson algorithm, the sample estimate of partial autocorrelation at lag 1 should be the same as that for autocorrelation at lag 1.

To conduct a test on autocorrelations or partial autocorrelations at certain lags, we can first construct a confidence interval using the VMB bootstrap and then examine whether the constructed interval contains 0.

[6] It means that adjacent points in the generated bootstrap series may be distant away from each other in the original series and hence are not likely to be correlated with each other despite that they are adjacent data points in the generated series and are supposed to be correlated with each other.

A noteworthy point regarding the VMB bootstrap is that a special jackknife procedure is needed for BCa intervals because the usual jackknife procedure for independent data cannot be directly applied to time series data. As a result, we used the jackknife procedure proposed by Künsch (1989) for BCa interval construction of the VMB bootstrap in this study. Similar to the VMB bootstrap, the Künsch's jackknife first constructs $T - b + 1$ consecutive blocks of length b. In the second step, however, instead of sampling T/b blocks from the constructed $T - b + 1$ blocks, this jackknife procedure iteratively deletes one block from the original series, yielding $T - b + 1$ new series. Then as in the VMB bootstrap, the Künsch's jackknife computes estimates of (partial) autocorrelations based on pairs of observations for each newly generated series.

Regarding the comparison between percentile and BCa intervals, because the VMB bootstrap works under the alternative, Fisher's z transformation fails to stabilize the standard error of the transformed statistic. Hence, BCa intervals are likely preferable to percentile intervals when combined with the VMB bootstrap.

Surrogate Data Method

While the bootstrap resamples from the original distribution that generates the observed data, the surrogate data method examines the null hypothesis by simulating the sampling distribution under the null and comparing the observed statistic with the confidence interval obtained from the estimated sampling distribution (Theiler et al., 1992). In the context of testing (partial) autocorrelations, the surrogate data method simulates the sampling distribution under the null by shuffling the order of the original series, given that the null hypothesis assumes that all lagged correlations are zero. The main difference between the surrogate data method and the moving block bootstrap is that the former matches the logic of hypothesis testing whereas the latter follows the logic of confidence interval construction.

A modification is needed when computing the bias constant z_0 for BCa intervals. The BCa method evaluates median bias by checking whether the original sample estimate (treated as the population value in the world of resampling) equals the median of resampling replications (treated as samples in the world of resampling). In the situation of testing (partial) autocorrelations, because the surrogate data method works under the null, what is wanted is the bias under the null. However, the observed series does not necessarily come from the null, and thus original sample estimates, i.e., the observed (partial) autocorrelations might not be substitutes for population values under the null. Therefore, we instead used population values under the null, i.e., zero. In particular, $\hat{z}_0$ is obtained by counting the proportion of replications with (partial) autocorrelation estimates less than zero (rather than sample estimates) and finding the corresponding percentile from the standard normal distribution.

Regarding the comparison between percentile and BCa intervals, because the surrogate data method works under the null, i.e., having zero population (partial) autocorrelations, Fisher's z transformation should satisfy the requirement of normalization and stabilization for percentile intervals. As a result, percentile intervals are likely to be comparable to BCa intervals. However, because population values are used instead of sample estimates when computing $\hat{z}_0$, BCa intervals might be inferior to percentile intervals. This possibility is explored in the simulation study.

A Simulation Study

In this section, we carried out a simulation study to evaluate and compare the performance of the six tests of autocorrelations and the five tests of partial autocorrelations. In the simulation, our main objective was to evaluate the performance of the tests under evaluation with complete data. Specifically, we studied the effect of nonnormality and time series length. Normal data and long series are preferred for accurate statistical inference. But these types of data are often uncommon in reality. Therefore, it is desirable to know which test is most tolerant of nonnormal data and short series.

Simulation Design

Data were simulated from a set of moving average (MA[3]) and autoregressive (AR[3]) models. Specifically, MA series were generated from MA models $y_t = b_1 v_{t-1} + b_2 v_{t-2} + b_3 v_{t-3} + v_t$ with MA weights b_1, b_2, b_3 determined by autocorrelations at the first three lags specified in Table 1[7]. To mimic real data, values for autocorrelations were chosen based on the average sample autocorrelations of the 146 daily moral affect series collected by Hardy et al. (2014), i.e., .266, .153, and .100 for the first three lags. Population partial

[7] Because the calculation of (partial) autocorrelations does not require explicit model fitting, instead of choosing values for AR or MA weights directly, we set up values for (partial) autocorrelations first.

autocorrelations were then computed using the `ARMAacf` function in R. Similarly, AR series were generated from AR models $y_t = a_1 y_{t-1} + a_2 y_{t-2} + a_3 y_{t-3} + v_t$ with AR weights a_1, a_2, a_3 determined by partial autocorrelations at the first three lags specified according to the average sample partial autocorrelations up to lag 3, i.e., .266, .040, and .023. Population autocorrelations were then computed accordingly. Model M0 with zero (partial) autocorrelations was used to explore Type I error rates. All models are stationary and invertible in the population.

Three potential influential factors were considered, as summarized in Table 2. First, we considered three time series lengths that are commonly seen in psychological studies: 50, 100 and 200. Second, distributions with heavy tails ($t_{4.5}$, excess kurtosis =12) and/or skewness (Gamma$_{(1,1)}$, skewness = 2 ; excess kurtosis = 6) were included to study the performance of various tests of (partial) autocorrelations with nonnormal data. We also considered three different levels of variance for the shock variable, from small to large: 0.5, 1, and 1.5.

In the simulation, we evaluated six tests of autocorrelations: the $1/T$ approximation method ($1/T$), the test based on the Bartlett's formula (Bartlett), the vectorized moving block bootstrap with percentile intervals (VMB) and with BCa intervals (VMB-BCa), the surrogate data method with percentile intervals (Surrogate) and with BCa intervals (Surrogate-BCa). For partial autocorrelations, the test based on the Bartlett's formula is unavailable, rendering the total number of tests to 5. All tests were used to test 1000 time series in every condition, and were evaluated in terms of Type I error rates and statistical power. The number of bootstrap replications / surrogate data sets was set to $B = 2000$. When implementing the VMB bootstrap, the block size b needs to be determined beforehand. In this study, we followed the recommendation by Künsch (1989) who suggested that an optimal block length $b \propto T^{\frac{1}{3}}$, and thus we used a block length of $b \approx T^{\frac{1}{3}}$. If the obtained b is not an integer, we rounded it to the nearest integer.

Results Regarding Autocorrelations

Figure 1 visualizes the empirical Type I error rates for different tests of autocorrelations with various time series lengths, shock variable distributions, and shock variable variances. Generally, the $1/T$ approximation method and the surrogate data method (with either percentile or BCa intervals) showed better control of Type I error than did the Bartlett's formula method and the VMB bootstrap (with either percentile or BCa intervals). When the time series length was short (e.g., $T = 50$) or/and data did not follow normal distributions, the $1/T$ approximation method was relatively conservative with underestimated Type I error rates and the surrogate data method performed better in this situation. More specifically, among the 27 conditions and the three parameters (autocorrelations at the first three lags) of interest, the $1/T$ approximation method produced Type I error rates significantly different from the nominal level 23 of 81(27×3) times, as indicated by the number of times the solid line with triangles falling outside the 95% confidence band of Type I error rates in Figure 1. In contrast, the number for the surrogate data method was 5 with percentile intervals and 11 with BCa intervals, both of which are substantially smaller than that of the $1/T$ approximation method.

Between the two types of tests with less satisfactory performance, the Bartlett's formula method was generally too conservative. As shown in Figure 1, this method produced conservative Type I error rates 75 of 81 times. In contrast, the VMB bootstrap was generally too liberal, especially when percentile intervals were used, which produced inflated Type I error rates 76 of 81 times. Results became better when BCa intervals were used, but still far from the acceptable range: inflated Type I error rates were observed 60 of 81 times.

Because the VMB bootstrap with either percentile or BCa intervals produced inflated type I error rates in most studied conditions, they were not included in the following power analysis.

Given that the $1/T$ approximation method is widely used, we contrasted it with the rest of tests except the VMB bootstrap with either percentile or BCa intervals because they showed inflated Type I error rates. As shown in Figure 2, the Bartlett's formula method was less powerful than the $1/T$ approximation method across all the studied conditions. The surrogate data method with BCa intervals seemed to perform slightly worse than the $1/T$ approximation method. In contrast, the surrogate data method with percentile intervals showed higher power than the $1/T$ approximation method across all the studied conditions.

A regression analysis was used to further analyze under which conditions the differences in power were substantial between the surrogate data method with percentile intervals and the $1/T$ approximation method. Regression results (see Table 3) indicated that differences in power were larger in conditions with normal or t distributed shock variables, or for shorter series.

Results Regarding Partial Autocorrelations

Figure 3 visualizes the empirical Type I error rates for the five tests of partial autocorrelations under conditions with various time series lengths, shock variable distributions, and shock variable variances. Generally, the pattern of results was similar to that of tests of autocorrelations. Specifically, the surrogate method with percentile intervals performed the best (Type I error rates significantly different from the intended level 7 out of 81 times); the $1/T$ approximation method and the surrogate data method with BCa intervals were comparable to each other and both of them were inferior to the surrogate method with percentile intervals (Type I error rates off the intended level 21 and 26 out of 81 times respectively); the VMB bootstrap was generally too liberal, regardless of being combined with percentile or BCa intervals (producing inflated Type I error rates 74 and 58 out of 81 times respectively).

Because the VMB bootstrap with either percentile or BCa intervals produced inflated type I error rates in most studied conditions, they were not included in the following power analysis.

Regarding the empirical power, again, we used the widely used $1/T$ approximation method as a benchmark. As shown in Figure 4, we found similar results as those for autocorrelations. In particular, the surrogate data method with percentile intervals showed higher power than the $1/T$ approximation method in most studied conditions, and when combined with BCa intervals, the surrogate data method showed lower empirical power.

Regression analysis on power differences between the surrogate data method with percentile intervals and the $1/T$ approximation method showed that differences were larger in conditions with normal or t distributed shock variables, shorter time series lengths, or larger effect sizes (see Table 4).

An R Package `pautocorr`

To ease the implementation of the six tests of autocorrelations and the five test of partial autocorrelations, we developed an R package `pautocorr`. This package is written for univariate time series. It calculates (partial) autocorrelation estimates, their standard errors or confidence intervals using the six (five) evaluated methods in this study. Currently, the package can only handle missing values using the available-case method[8]. Below, we introduce the use of this

[8] When computing (partial) autocorrelation estimates, pairs of observations with missing values are removed.

package.

Included in Appendix is the example R code illustrating the use of the package. To use the package within R, first to install the package using the command `install.packages("pautocorr",repos="http://R-Forge.R-project.org")` (Line 2) and then load it using the command `library(pautocorr)` (Line 3). The code in Line 4 and Line 5 loads the data set `daily.moral` into R and extracts the moral affect series from the data set. Often, observed series are not stationary and so are our example data. The built-in function `decompose` is helpful in making observed series stationary (Line 7). To use function `decompose`, researchers have to redefine the series in the format of time series data using function `ts` (Line 6). Function `pautocorr.test` is the main function from the package `pautocorr` to conduct the six tests of autocorrelations and the five tests of partial autocorrelations. The `lagmax` argument in this function specifies a maximum time lag for (partial) autocorrelations to be tested. The argument `alpha` sets the significance level. The number of resampling replications can be changed via the argument `B` and the upper limit for the summed terms in the Bartlett's formula can be specified through the argument `L`. Numerical results will be returned if the argument `print` is set to `TRUE`. To visualize results, use the function `plot`. Normally, a graph with four plots visualizing the results of the tests under evaluation will be generated. An example is given in the next section *Real Data Illustration*.

Real Data Illustration

We now demonstrate the use of the six tests of autocorrelations and the five tests of partial autocorrelations using an empirical study. The data used for illustration included 50 days of self-ratings on daily moral affect (assessed by the sum of scores for empathy, gratitude and forgiveness emotions) from one participant. The series was extracted from a data set of a large study by Hardy et al. (2014), which investigated the relationships among daily religious involvement, spirituality and moral emotions. The current illustration focuses on lagged effects of daily moral affect.

The history plot (see the top left graph in Figure 5) suggested that the original series contained noticeable long term change which made the series nonstationary. To make the series stationary, we removed the trend component that reflected long term change using the `decompose` function in R. We then tested the stationarity of the residual series through the augmented Dickey-Fuller unit root test (Dickey & Fuller, 1979, 1981) using the `ur.df` function in the `urca` package in R. Results confirmed that the residual series was stationary ($t = -5.55$, $p < 0.01$). The history plot of the residual series echoed the results (see the top right graph in Figure 5).

As shown by the ACF and PACF plots (see the two middle graphs in Figure 5), the results of the $1/T$ approximation method and the surrogate data method with BCa intervals suggested that the underlying model might be an MA(3) or AR(3) model since only autocorrelation and partial autocorrelation at lag 3 were significant. One could subjectively interpret these results as the ACF cut off after lag 3 and the PACF tailed off or the other way around. The results of the surrogate data method with percentile intervals however showed that it was more likely to be an AR(3) model given that autocorrelation at lag 5 was also significant.

The bottom two plots showed that the VMB bootstrap with BCa intervals and the Bartlett's formula method found similar results as did the $1/T$ approximation method and the surrogate data method with BCa intervals. Just as in the simulation study, the VMB bootstrap with percentile intervals rejected the null most frequently. Based on simulation results, we do not recommend using the Bartlett's formula method or the VMB bootstrap.

To sum up, the illustration here paralleled the results of the simulation study. In particular, the surrogate data method with percentile intervals had higher power than the $1/T$ approximation method and the surrogate data method with BCa intervals. And this difference might lead to different conclusions in model order identification in real data analysis.

Discussion

In this study, we compared six tests of autocorrelations and five tests of partial autocorrelations using a simulation study and a real data example. Generally, the surrogate data method with percentile intervals outperformed the other tests in terms of better control of Type I error and higher statistical power. This finding is generally in accordance with the study on tests of correlations for complete but independent data (Lee & Rodgers, 1998). Additionally, the use of BCa intervals resulted in substantial decrease in power, suggesting not to use BCa intervals together with the surrogate data method. We believe that this is related to the inaccurate estimation of the median bias. With inaccurate estimates of median bias, bias correction is of course inaccurate. Consequently, BCa intervals may perform less satisfactorily.

The advantage of the surrogate data method over the $1/T$ approximation method decreased as time series length increased, which is as expected given its roots in large sample behaviors. When time series length is sufficiently large (e.g., $T \geq 200$ in this study), the $1/T$ approximation method can be used due to its simplicity.

Based on the simulation study, the test based on the Bartlett's formula is generally not recommended for testing (partial) autocorrelations because of its conservative Type I error rates and low statistical power. The two VMB bootstrap methods are not suitable for testing (partial) autocorrelations as well, as suggested by the inflated Type I error rates in the simulation study. We believe that the main reason for the unsatisfactory performance of the VMB bootstrap is that for statistics such as sample (partial) autocorrelations, the sampling distribution under the null has a different shape compared to that under the alternative. Consequently, the chances of estimated (partial) autocorrelations falling into the 95% confidence intervals under the null do not equal the chances of H_0 value, that is 0 in our situation, falling into the 95% confidence intervals under the alternative. This finding is somewhat contradictory to the study on bootstrap standard error estimates in dynamic factor analysis by Zhang & Browne (2010). In that study, Type I error rates of the VMB bootstrap were inflated but still within the acceptable range. One possible explanation is that in the simulation study, Zhang & Browne (2010) used normal shock variables and error terms only. In the current study, we considered nonnormal shock variables as well. In this study, the control of Type I error of the VMB bootstrap was worse in conditions with nonnormal shock variables. Additionally, this study showed that compared to percentile intervals, the VMB bootstrap worked better with BCa intervals. In conclusion, we recommend using the surrogate data method with percentile intervals to test the hypothesis whether (partial) autocorrelations at certain lags are zero or to build AR or MA models, given that it shows better control on Type I error and higher power.

The current study has several limitations. First, we did not consider model selection techniques such as the information criterion methods (e.g., AIC and BIC indices) and the Hannan-Rissanen method (Hannan & Rissanen, 1982). Theoretically, the use of information criteria requires explicitly fitting a model, which is avoided by the methods evaluated in this study. Additionally, the “model” estimated while applying information criteria not only specifies the structural model but also the distributional model. Because normality is often assumed, for nonnormal data with a small sample size, complicated modification is often necessary (e.g., Burnham & Anderson, 2002; Konishi & Kitagawa, 2008). Given these two considerations, we decided to leave information criterion methods for future studies so that we could focus on the objectives of the current study. The Hannan-Rissanen method avoids multiple model estimation of the information criterion methods by comparing models through linear regression calculations. However, this method and its improved version (e.g., Poskitt, 1987) still rely on minimizing information criteria such as BIC or other indices. Their robustness to nonnormality in small sample size conditions is unwarranted. Despite those considerations, we believe that the information criterion methods and the Hannan-Rissanen method are competitive alternatives. Future research may systematically compare those model selection techniques with the studied methods here, especially for nonnormal data with small sample sizes.

Second, we did not consider other tests of autocorrelations such as the Box-Pierce test, the Ljung-Box test and their extended versions (e.g., Box & Pierce, 1970; Ljung & Box, 1978; Horowitz et al., 2006; Lobato et al., 2001, 2002). The Box-Pierce and Ljung-Box tests examine whether the autocorrelations up to lag l are all zero. Hence, both tests are more suitable for testing the overall randomness of a time series. If, however, testing the significance of the autocorrelation at lag l is of interest, e.g., in the case of model order identification, it is generally difficult to make inference based on the results of these two tests. For example, suppose the observed series is generated from an AR(3) model. What can be inferred from the Box-Pierce and Ljung-Box tests is that the observed series is or is not serially correlated. Researchers still need to rely on other inferential techniques. Therefore, in this study, we did not examine the performance of the Box-Pierce test, the Ljung-Box test and their extended versions.

Finally, we did not consider the missing data problem, which is a common problem in time series data analysis. After years of study, it is clear that it is crucial to distinguish between missing mechanisms. According to Little & Rubin (2002), there are three types of missing mechanisms: missing completely at random (MCAR, the missing mechanism is a complete random process), missing at random (MAR, the missingness can be fully predicted by observed values), and missing not at random (MNAR, the missingness is related to unobserved values).

When the missing mechanism is MCAR, observed data points can be viewed as a subsample of the original sample. Consequently, one can expect that the studied tests of (partial) autocorrelations combined with the available-case method (e.g., see Little & Rubin, 2002, pp.53) for missing values would show similar patterns of results as those in this study. Other missing mechanisms such as MAR or MNAR requires sophisticated methods to handle the missing data problem, e.g., the imputation method or the EM algorithm with auxiliary variables (e.g., see Little & Rubin, 2014). In time series analysis, the nature of time dependence makes it highly challenging to deal with missing data at the stage of model order identification. This is because the process of model order identification and missing data correction may influence each other and become entangled with each other. We believe that there is no easy solution to the missing data problem. Future research may systematically study this issue.

In sum, in this study, we have evaluated the finite sample performance of the six tests of (partial) autocorrelations for both normal and nonnormal series. Based on simulation results, we recommend the surrogate data method with percentile intervals given its robustedness to nonnormality and high statistical power. Tests of (partial) autocorrelations are useful for model order identification for AR or MA models. In real data analysis, model order identification is a challenging task given the reality of small sample size, nonnormality and other issues. We believe the surrogate data method with percentile intervals is useful for this task.

References

Bartlett, M. S. (1955). *An introduction to stochastic processes.* Cambridge: Cambridge University Press.

Borckardt, J. J., Nash, M. R., Murphy, M. D., Moore, M., Shaw, D., & O'Neil, P. (2008). Clinical practice as natural laboratory for psychotherapy research: A guide to case-based time series analysis. *American Psychologist*, *63*, 77-95. doi: 10.1037/0003-066X.63.2.77.

Box, G. E. P., & Pierce, D. A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association*, *65*(332), 1509-1526. doi: 10.1080/01621459.1970.10481180

Brockwell, P. J., & Davis, R. A. (2002). *Introduction to time series and forecasting* (2nd ed.). New York: Springer.

Browne, M. W., & Nesselroade, J. R. (2005). Representing psychological processes with dynamic factor models: Some promising uses and extensions of ARMA time series models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Advances in psychometircs: A festschrift for Roderick P. McDonald* (p. 415-452). Mahwah: Erlbaum.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). Springer-Verlag.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, *74*(366a), 427-431. doi: 10.1080/01621459.1979.10482531

Dickey, D. A., & Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica*, *49*(4), 1057-1072. doi: 10.2307/1912517

du Toit, S. H. C., & Browne, M. W. (2007). Structural equation modeling of multivariate time series. *Multivariate Behavioral Research*, *42*(1), 67-101. doi: 10.1080/00273170701340953

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*(397), 171-185. doi: 10.2307/2289144

Efron, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, *89*(426), 463-475. doi: 10.1080/01621459.1994.10476768

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Chapman and Hall/CRC.

Ferrer, E., & Zhang, G. (2009). The series models for examining psychological processes: Applications and new developments. In R. E. Millsap & A. Madeu-Olivares (Eds.), *Handbook of quantitative methods in psychology* (p. 637-657). New Bury Park: Sage.

Frangos, C. C., & Schucany, W. R. (1990). Jackknife estimation of the bootstrap acceleration constant. *Computational Statistics & Data Analysis*, *9*(3), 271–281. doi: 10.1016/0167-9473(90)90109-u

Hannan, E. J., & Rissanen, J. (1982). Recursive estimation of mixed autoregressive-moving average order. *Biometrika*, *69*(1), 81-94. doi: 10.1093/biomet/69.1.81

Hardy, S. A., Zhang, Z., Skalski, J. E., & Melling, B. S. (2014). Daily religious involvement, spirituality, and moroal emotions. *Psychology of Religion and Spirituality*, *6*(4), 338-348. doi: 10.1037/a0037293

Hawkins, D. L. (1989). Using U statistics to derive the asymptotic distribution of Fisher's Z statistic. *The American Statistician*, *43*(4), 235-237. doi: 10.2307/2685369

Horowitz, J. L., Lobato, I., Nankervis, J. C., & Savin, N. (2006). Bootstrapping the Box–Pierce Q test: A robust test of uncorrelatedness. *Journal of Econometrics*, *133*(2), 841-862. doi: 10.1016/j.jeconom.2005.06.014

Konishi, S., & Kitagawa, G. (2008). *Information criteria and statistical modeling*. Springer.

Künsch, H. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, *17*(3), 1217-1241. doi: 10.1214/aos/1176347265

Lee, W.-C., & Rodgers, J. L. (1998). Bootstrapping correlation coefficients using univariate and bivariate sampling. *Psychological Methods*, *3*(1), 91-103. doi: 10.1037/1082-989X.3.1.91

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, New Jersey: John Wiley & Sons.

Little, R. J. A., & Rubin, D. B. (2014). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, *65*(2), 297-303. doi: 10.1093/biomet/65.2.297

Lobato, I., Nankervis, J. C., & Savin, N. (2002). Testing for zero autocorrelation in the presence of statistical dependence. *Econometric Theory*, *18*(3), 730-743. doi: 10.1017/s0266466602183083

Lobato, I., Nankervis, J. C., & Savin, N. E. (2001). Testing for autocorrelation using a modified Box-Pierce Q test. *International Economic Review*, *42*(1), 187-205. doi: 10.1111/1468-2354.00106

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156-166. doi: 10.1037/0033-2909.105.1.156

Poskitt, D. S. (1987). A modified Hannan-Rissanen strategy for mixed autoregressive-moving average order determination. *Biometrika*, *74*(4), 781-790. doi: 10.2307/2336472

Shumway, R. H., & Stoffer, D. S. (2006). *Time series analysis and its applications with R examples* (2nd ed.). New York: Springer.

Theiler, J., Eubank, S., Longtin, A., Galdrikan, B., & Farmer, J. D. (1992). Testing for nonlinearity in time series: The method of surrogate data. *Physica D: Nonlinear Phenomena*, *58*, 77-94. doi: 10.1016/0167-2789(92)90102-S

Zhang, G., & Browne, M. W. (2010). Bootstrap standard error estimates in dynamic factor analysis. *Multivariate Behavioral Research*, *45*, 453-482. doi: 10.1080/00273171.2010.483375

Appendix: Example R Code

Example R code illustrating the use of the customized R functions for the six tests of autocorrelations

```
##      Input   ##
install.packages('pautocorr',repos="http://R-Forge.R-project.org")
library(pautocorr)
data('daily.moral')
d = daily.moral[,'daily.moral']
ts.d = ts(data = d,frequency = 7)
d.adj = na.omit(d - decompose(ts.d)$trend)
res = pautocorr.test(d.adj,lagmax=15,alpha=0.05,B=2000,print = TRUE)
plot(res)
```

Table 1

Parameter / population values used in the simulation.

		MA Weights			Autocorrelations			Partial Autocorrelations		
Models		b_1	b_2	b_3	ρ_1	ρ_2	ρ_3	ϕ_1	ϕ_2	ϕ_3
	M0	0	0	0	0	0	0	0	0	0
MA(3)	M1	.187	.095	.052	.200	.100	.050	.200	.063	.020
	M2	.382	.194	.120	.400	.200	.100	.400	.048	.006
		AR Weights			Autocorrelations			Partial Autocorrelations		
		a_1	a_2	a_3	ρ_1	ρ_2	ρ_3	ϕ_1	ϕ_2	ϕ_3
AR(3)	M3	.175	.091	.050	.200	.136	.092	.200	.100	.050
	M4	.300	.168	.100	.400	.328	.266	.400	.200	.100

Note: Values for autocorrelations of MA models and partial autocorrelations of AR models were chosen according to analysis results of real series. MA, AR weights, partial autocorrelations of MA models and autocorrelations of AR models were computed according to the theoretical relationship among MA (or AR) weights, autocorrelations and partial autocorrelations of MA (or AR) series.

Table 2

Influential factors considered in Study 1.

Factors	Description
Time series length (T)	50,100,200
Population distribution of $v_1 - v_T$	$N\left(0, \sigma_v^2\right)$, $t_{4.5}\left(0, \sigma_v^2\right)$, $\Gamma_{(1,1)}\left(0, \sigma_v^2\right)$
variance (σ_v^2) of $v_1 - v_T$	0.5, 1, 1.5

Note: $t_{4.5}\left(0, \sigma_v^2\right)$ and $\Gamma_{(1,1)}\left(0, \sigma_v^2\right)$ are the rescaled t and Gamma distributions with means and variances specified inside the parenthesis respectively.

Table 3

Regression analysis on power differences between the surrogate data method with percentile intervals and the $1/T$ approximation method for tests of autocorrelations.

Coef.	Est.	Std. Err.	$p-$Value
Intercept	0.019	0.003	<.001
Normal	0.014	0.002	<.001
$t_{4.5}$	0.018	0.002	<.001
$T = 50$	0.019	0.002	<.001
$T = 200$	-0.016	0.002	<.001
$\sigma_v^2 = 1$	0.000	0.002	0.837
$\sigma_v^2 = 1.5$	0.000	0.002	0.883
Effect Size	0.012	0.008	0.124

Table 4

Regression analysis on power differences between the surrogate data method with percentile intervals and the $1/T$ approximation method for tests of partial autocorrelations.

Coef.	Est.	Std. Err.	$p-$Value
Intercept	0.017	0.004	<.001
Normal	0.011	0.003	<.001
$t_{4.5}$	0.013	0.003	<.001
$T = 50$	0.013	0.003	<.001
$T = 200$	-0.011	0.003	<.001
$\sigma_v^2 = 1$	0.001	0.003	0.809
$\sigma_v^2 = 1.5$	0.001	0.003	0.951
Effect Size	0.037	0.010	<.001

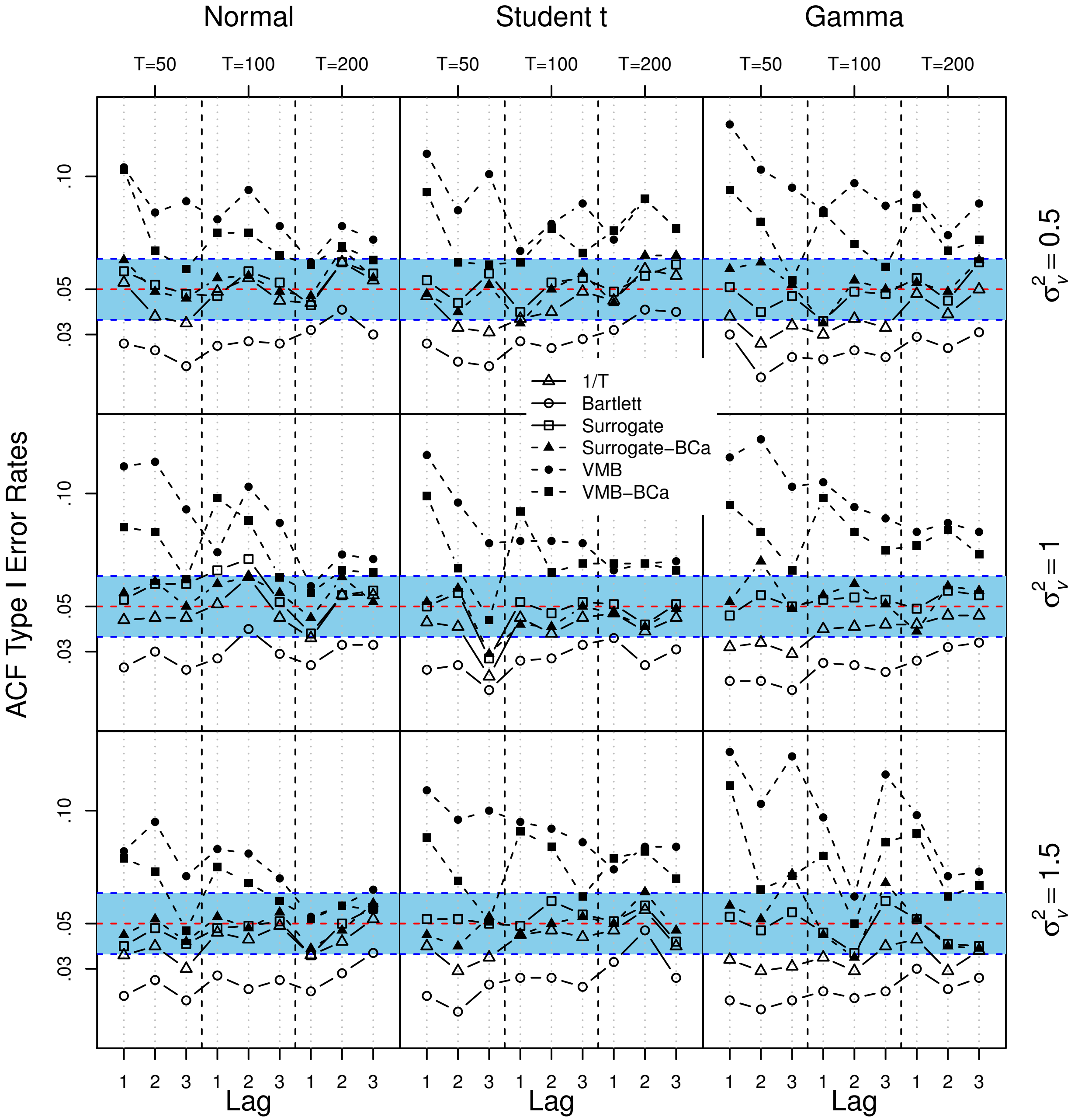

Figure 1. Empirical Type I error rates for the six tests of autocorrelations (the $1/T$ approximation method [$1/T$], the Bartlett's formula method [Bartlett], the surrogate data method with percentile [Surrogate] and BCa intervals [Surrogate-BCa], and the vectorized moving block bootstrap with percentile [VMB] and BCa intervals [VMB-BCa]) as a function of various time series lengths, shock variable distributions and shock variable variances. Note: The 95% confidence band for Type I error rates ($.05 \pm 1.96\sqrt{\frac{0.05\times0.95}{1000}} \approx [.036, .064]$) was shaded in blue.

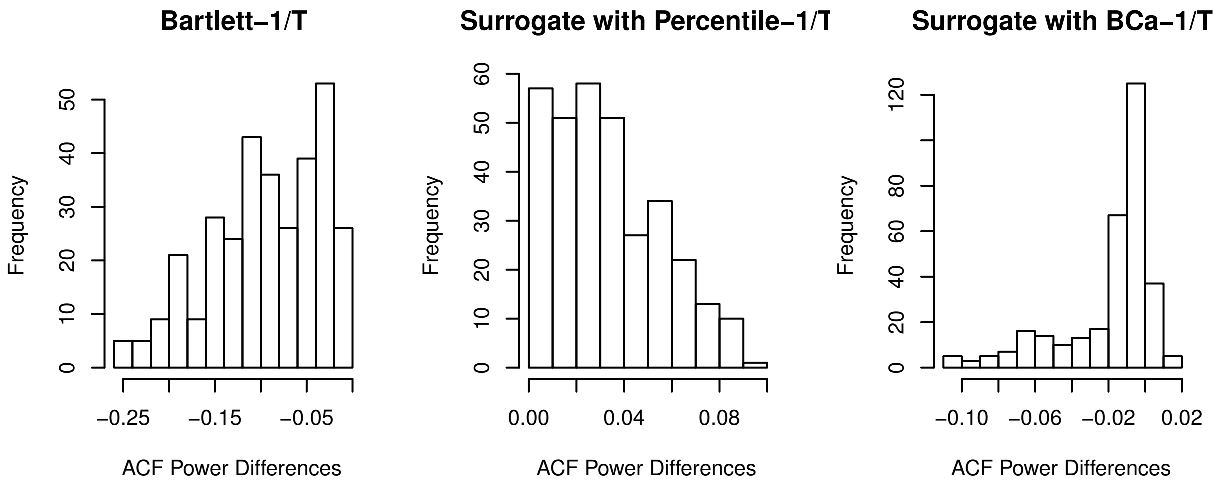

Figure 2. Contrast the power of the $1/T$ approximation method with that of the remaining tests of autocorrelations except for the moving block bootstrap. “BCa” is short for bias-corrected and accelerated intervals.

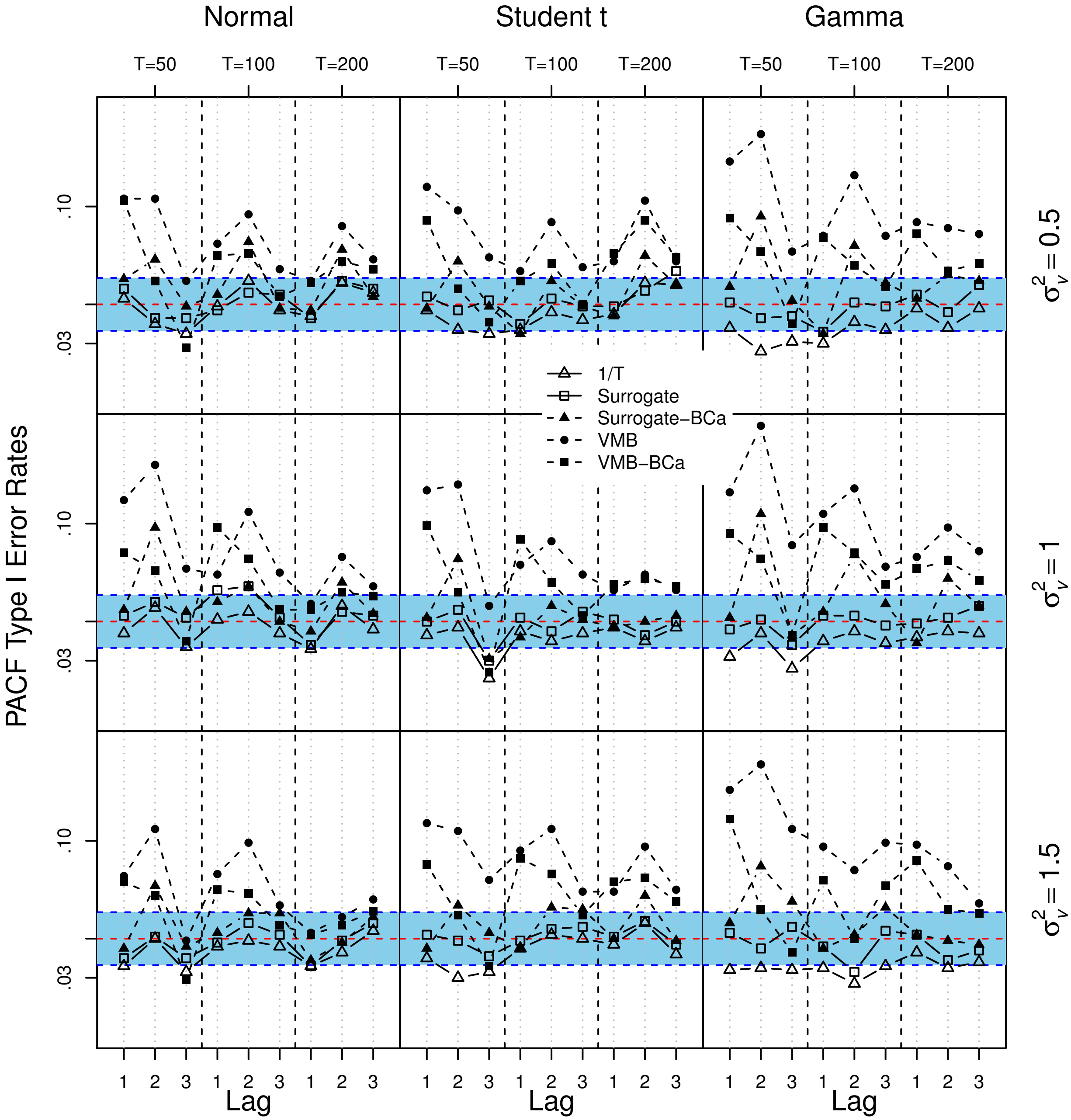

Figure 3. Empirical Type I error rates for the five tests of partial autocorrelations (the $1/T$ approximation method [$1/T$], the surrogate data method with percentile [Surrogate] and BCa intervals [Surrogate-BCa], and the vectorized moving block bootstrap with percentile [VMB] and BCa intervals [VMB-BCa]) as a function of various time series lengths, shock variable distributions and shock variable variances. Note: The 95% confidence band for Type I error rates ($.05 \pm 1.96\sqrt{\frac{0.05\times0.95}{1000}} \approx [.036, .064]$) was shaded in blue.

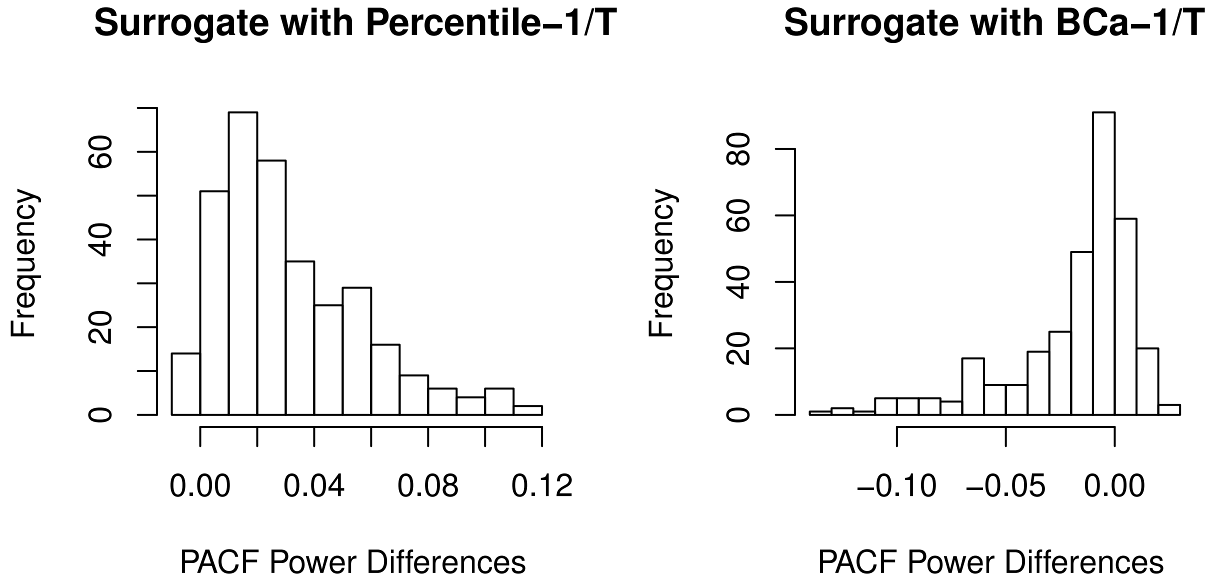

Figure 4. Contrast the power of the $1/T$ approximation method with that of the surrogate data method with percentile or BCa intervals for partial autocorrelation testing. “BCa” is short for bias-corrected and accelerated intervals.

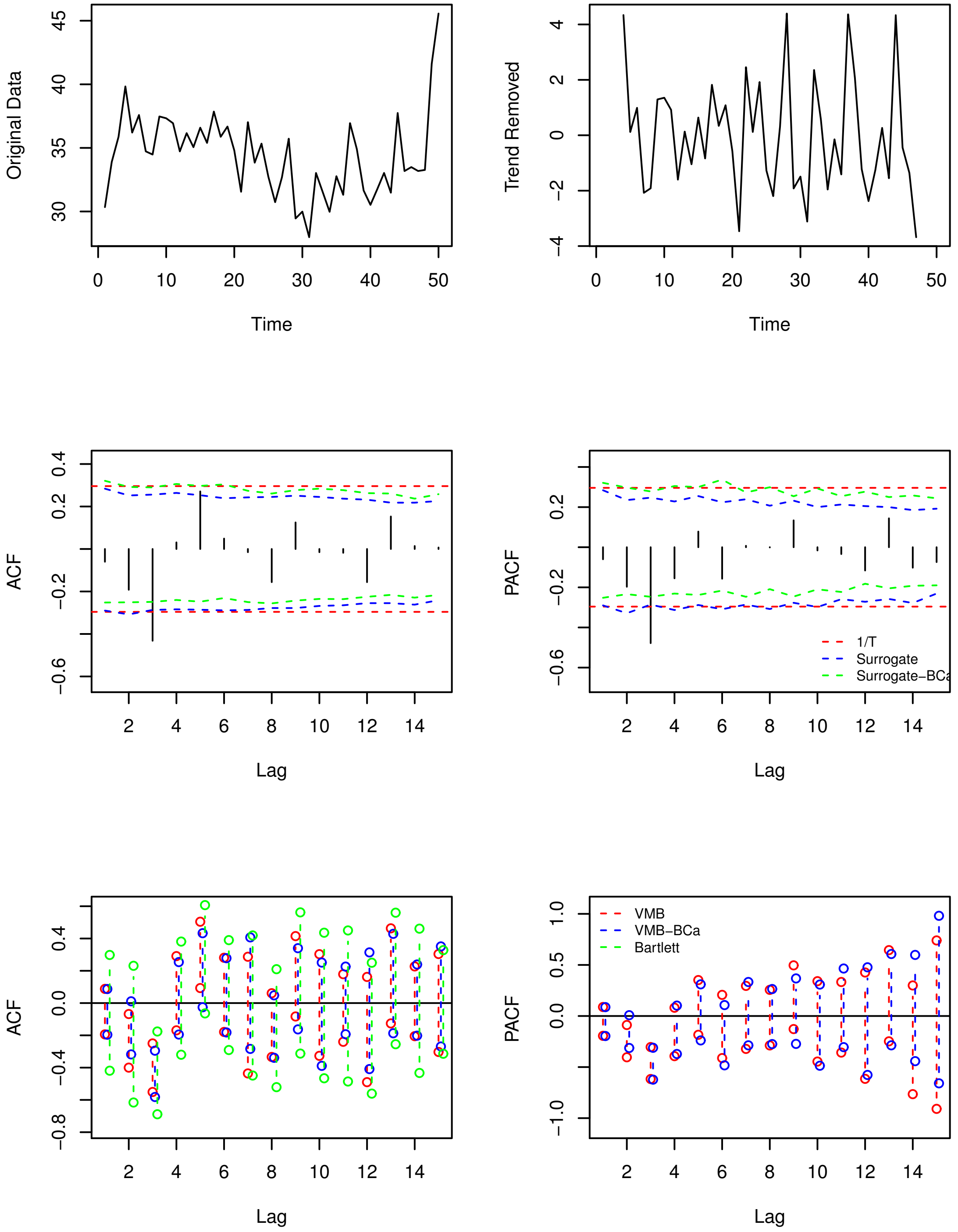

Figure 5. Test the ACF and the PACF of a daily moral affect series. “$1/T$”: the $1/T$ approximation method; “Bartlett”: the Bartlett’s formula method; “Surrogate” and “Surrogate-BCa”: the surrogate data method with percentile and BCa intervals respectively; and “VMB” and “VMB-BCa”: the vectorized moving block bootstrap with percentile and BCa intervals respectively.