Supplemental Material for "*Testing the White Noise Hypothesis of Stock Returns*"

Jonathan B. Hill* – University of North Carolina
Kaiji Motegi† – Kobe University

This draft: August 3, 2018

*Department of Economics, University of North Carolina at Chapel Hill. E-mail: jbhill@email.unc.edu
†*Corresponding author.* Graduate School of Economics, Kobe University. E-mail: motegi@econ.kobe-u.ac.jp

Contents

List of Tables

List of Figures

1 Introduction

In the main paper Hill and Motegi (2018), we perform a rolling window analysis of the autocorrelation structure of stock returns. Confidence bands for sample autocorrelations and a Cramér-von Mises [CvM] white noise test statistic are constructed with Shao's (2011) blockwise wild bootstrap. We observe a perverse phenomenon that the confidence bands go up and down periodically, and the span of the periodical movement is equal to block size b_n. In what follows, we give a detailed explanation for the periodicity, and propose randomizing the block size across bootstrap samples and rolling windows as a solution, similar to the stationary bootstrap in Politis and Romano (1994). Periodicity of bootstrapped samples is well known to occur in the block bootstrap for time series data. See Politis and Romano (1994) and Lahiri (1999).

In Section 2, we review Andrews and Ploberger's (1996) [AP] sup-LM test statistic and Hill and Motegi's (2017) max-correlation test statistic. Based on the theoretical and numerical analysis of Hill and Motegi (2017), the AP, max-correlation, and CvM statistics attain better size and power than other white noise test statistics. The main paper Hill and Motegi (2018) focuses on the CvM statistic to save space, while this supplemental material covers all of the three test statistics.

In Section 3 we review the procedure of the blockwise wild bootstrap and show that periodicity appears even for iid data. The same issue arises in sample autocorrelations and any function of theirs, including critical values and p-values for the max-correlation, AP, CvM tests. In Section 4 we provide a detailed explanation for the periodicity. In Section 5 we show that randomizing a block size removes periodicity as desired. In Section 6 we perform Monte Carlo simulations in a full sample environment in order to show that the blockwise wild bootstrap with randomized block size yields decent size and power in finite sample. In Section 7 we report empirical results on stock returns that are omitted in the main paper.

2 Alternative White Noise Test Statistics

We begin with a brief review of the notation in the main paper. Let $\{r_t\}$ be a time series of interest (e.g. stock return). Define population mean $\mu = E[r_t]$, variance $\gamma(0) = E[(r_t - \mu)^2]$, autocovariance $\gamma(h) = E[(r_t - \mu)(r_{t-h} - \mu)]$, and autocorrelation $\rho(h) = \gamma(h)/\gamma(0)$ for $h \geq 1$. We wish to test the white noise hypothesis:

$$H_0 : \rho(h) = 0 \text{ for all } h \geq 1 \quad \text{against} \quad H_1 : \rho(h) \neq 0 \text{ for some } h \geq 1.$$

Write sample mean $\hat{\mu}_n = (1/n)\sum_{t=1}^n r_t$, variance $\hat{\gamma}_n(0) = (1/n)\sum_{t=1}^n (r_t - \hat{\mu}_n)^2$, autocovariance $\hat{\gamma}_n(h) = (1/n)\sum_{t=h+1}^n (r_t - \hat{\mu}_n)(r_{t-h} - \hat{\mu}_n)$, and autocorrelation $\hat{\rho}_n(h) = \hat{\gamma}_n(h)/\hat{\gamma}_n(0)$ for $h \geq 1$. In

order to ensure a valid white noise test and therefore capture all serial correlations asymptotically, we formulate test statistics based on the serial correlation sequence $\{\hat{\rho}_n(h)\}_{h=1}^{\mathcal{L}_n}$ with sample-size dependent lag length $\mathcal{L}_n \to \infty$ as $n \to \infty$.

The sup-LM test statistic proposed by Andrews and Ploberger (1996) has an equivalent representation (see Nankervis and Savin, 2010):

$$\mathcal{AP}_n = \sup_{\lambda \in \Lambda} \left\{ n(1-\lambda^2) \left(\sum_{h=1}^{\mathcal{L}_n} \lambda^{h-1} \hat{\rho}_n(h) \right)^2 \right\} \quad \text{where} \quad \mathcal{L}_n = n-1,$$

where Λ is a compact subset of $(-1, 1)$. The latter ensures a non-degenerate test that obtains, under suitable regularity conditions, an asymptotic power of one when there is serial correlation at some horizon.

Andrews and Ploberger (1996) use $\mathcal{L}_n = n-1$ for computing the test statistic, but truncate a Gaussian series that arises in the limit distribution in order to simulate critical values. Nankervis and Savin (2010, 2012) generalize the sup-LM test to account for data dependence, and truncate the maximum lag both during computation (hence $\mathcal{L}_n < n-1$), and for the sake of simulating critical values. The truncated value used, however, does not satisfy $\mathcal{L}_n \to \infty$ as $n \to \infty$, hence their version of the test is inconsistent (i.e. it does not achieve asymptotic power 1 when the null hypothesis is false). To control for possible dependence under the null, and allow for a better approximation of the small sample distribution, we bootstrap the test with Shao's (2011) blockwise wild bootstrap.

The max-correlation test statistic proposed by Hill and Motegi (2017) is

$$\hat{\mathcal{T}}_n = \sqrt{n} \max_{1 \le h \le \mathcal{L}_n} |\hat{\rho}_n(h)|.$$

We require $\mathcal{L}_n \to \infty$ and $\mathcal{L}_n / n \to 0$ so that $\hat{\rho}_n(h) \xrightarrow{p} \rho(h)$ for each $h \in [1, \mathcal{L}_n]$. If the sequence of serial correlations is asymptotically iid Gaussian under the null, then the limit law of a suitably normalized $\hat{\mathcal{T}}_n$ under the null is a Type I extreme value, or Gumbel, distribution. That result extends to dependent data under the null (see Xiao and Wu, 2014, for theory and references). The non-standard limit law can be bootstrapped, as in Xiao and Wu (2014), although they do not prove the asymptotic validity of their double blocks-of-blocks bootstrap. Hill and Motegi (2017) sidestep an extreme value theoretic argument, and directly prove that Shao's (2011) blockwise wild bootstrap is valid without requiring the null limit law of the max-correlation. They also sidestep Gaussian approximation theory exploited in, amongst others, Chernozhukov, Chetverikov, and Kato

(2013), allowing for a very general setting and filtered data. See Hill and Motegi (2017) for a broad literature review and discussion.

Hill and Motegi (2017) find that the AP, max-correlation, and CvM tests, each assisted by the blockwise wild bootstrap, have comparable size and power in finite samples. When lag length $\mathcal{L}_n$ is large relative to sample size, the max-correlation test dominates the others in terms of size and power.[1]

3 Blockwise Wild Bootstrap in Rolling Windows

3.1 Bootstrap Procedure

We first review how to construct a 95% confidence band for sample autocorrelation at lag h, $\hat{\rho}_n(h)$, using the blockwise wild bootstrap. The algorithm is as follows. Set a block size b_n such that $1 \leq b_n < n$. Denote the blocks by $\mathcal{B}_s = \{(s-1)b_n + 1, \ldots, sb_n\}$ with $s = 1, \ldots, n/b_n$. Assume for simplicity that the number of blocks n/b_n is an integer. Generate iid random numbers $\{\xi_1, \ldots, \xi_{n/b_n}\}$ with $E[\xi_s] = 0$, $E[\xi_s^2] = 1$, and $E[\xi_s^4] < \infty$. In simulation experiments we draw ξ_s from the standard normal distribution. Define an auxiliary variable $\omega_t = \xi_s$ if $t \in \mathcal{B}_s$. Compute

$$\hat{\rho}_n^{(bwb)}(h) = \frac{1}{\hat{\gamma}_n(0)} \frac{1}{n} \sum_{t=h+1}^{n} \omega_t \left\{ \tilde{r}_t \tilde{r}_{t-h} - \hat{\gamma}_n(h) \right\}. \tag{1}$$

Repeat M times, resulting in $\{\hat{\rho}_{n,i}^{(bwb)}(h)\}_{i=1}^{M}$. The rest of the procedures is the same as the standard wild bootstrap.

The wild bootstrap can be interpreted as the blockwise wild bootstrap with $b_n = 1$ up to the centering of cross term. See Wu (1986) and Liu (1988) for seminal papers on the wild bootstrap, and see Hansen (1996). Asymptotic theory of the wild bootstrap is based on iid or mds properties. Hence the wild bootstrap is not a proper approach when, as in the present paper, it is of interest to test for serial uncorrelatedness strictly.

3.2 Periodicity of Rolling Window Confidence Bands

In the main paper, we compute rolling window sample autocorrelations of stock returns at lag 1 and their 95% confidence bands using the blockwise wild bootstrap. The resulting bands

[1] Other well-known test statistics include a standardized periodogram statistic of Hong (1996), which is effectively a standardized portmanteau statistic with a maximum lag $\mathcal{L}_n = n - 1$. The test statistic has a standard normal limit under the null, but Hill and Motegi (2017) show that the asymptotic test yields large size distortions. They also show that a bootstrap version, which is arithmetically equivalent to a bootstrapped portmanteau test, is often too conservative compared with the tests used in the present study. We therefore do not include Hong's (1996) test here.

have a periodic pattern that goes up and down rhythmically in every b_n windows. In this section we provide a small scale simulation experiment in order to show that the problem is endemic: it applies to any time series, irrespective of serial dependence.

Example 1 *We simulate iid standard normal* $\{y_t\}_{t=1}^T$ *with* $T = 3000$*, which resembles the sample size in the main paper. The 95% confidence bands are constructed by the wild bootstrap and blockwise wild bootstrap (cf. Section 3.1). Window sizes are* $n \in \{240, 480, 720\}$ *as in the main paper. For the blockwise wild bootstrap, block size is* $b_n = c\sqrt{n}$ *with* $c \in \{0.5, 1, 2\}$*, hence* $b_n \in \{7, 15, 30\}$ *for* $n = 240$*;* $b_n \in \{10, 21, 43\}$ *for* $n = 480$*;* $b_n \in \{13, 26, 53\}$ *for* $n = 720$*. The number of bootstrap iterations is* $M = 5000$ *for each window.*

See Figures 1-3 for results. The blockwise wild bootstrap produces zigzag confidence bands over the subsample windows for any block size, contrary to the true iid data generating structure. Furthermore, the span of periodic behavior is always equal to block size b_n*. The wild bootstrap does not generate any periodicity since it is interpreted as the blockwise wild bootstrap with* $b_n = 1$*.*

4 Reason for the Periodicity of Bootstrapped Confidence Bands

In this section we explain why the blockwise wild bootstrap produces periodic confidence bands, using an illustrative example. Consider a simple setting where the entire sample is $\{y_1, y_2, \ldots, y_{71}\}$, window size is $n = 60$, and block size is $b_n = 3$. Hence we have $71 - 60 + 1 = 12$ windows and each window contains $n/b_n = 20$ blocks.

The first window has

$$[y_1, y_2, y_3, \ldots, y_{59}, y_{60}]'$$

and we compute cross terms with lag 1:

$$[\underbrace{y_1y_2,\ y_2y_3}_{\mathcal{B}_1},\ \underbrace{y_3y_4,\ y_4y_5,\ y_5y_6}_{\mathcal{B}_2},\ \underbrace{y_6y_7,\ y_7y_8,\ y_8y_9}_{\mathcal{B}_3},\ \ldots,\ \underbrace{y_{57}y_{58},\ y_{58}y_{59},\ y_{59}y_{60}}_{\mathcal{B}_{20}}]',$$

where $\mathcal{B}_s$ signifies the s-th block. We then generate $\xi_1^{(1)}, \ldots, \xi_{20}^{(1)} \overset{i.i.d.}{\sim} N(0,1)$ and compute

$$W_1 = [\underbrace{\xi_1^{(1)}y_1y_2,\ \xi_1^{(1)}y_2y_3}_{\mathcal{B}_1},\ \underbrace{\xi_2^{(1)}y_3y_4,\ \xi_2^{(1)}y_4y_5,\ \xi_2^{(1)}y_5y_6}_{\mathcal{B}_2},\ \underbrace{\xi_3^{(1)}y_6y_7,\ \xi_3^{(1)}y_7y_8,\ \xi_3^{(1)}y_8y_9}_{\mathcal{B}_3},$$
$$\ldots,\ \underbrace{\xi_{20}^{(1)}y_{57}y_{58},\ \xi_{20}^{(1)}y_{58}y_{59},\ \xi_{20}^{(1)}y_{59}y_{60}}_{\mathcal{B}_{20}}]',$$

where superscript (1) on ξ's signifies that they are random numbers drawn for window 1. The sample mean of W_1 forms a bootstrapped autocorrelation $\hat{\rho}_n^{(bwb)}(1)$ up to mean centering and scaling.

A crucial feature that generates periodicity is that W_1, W_4, W_7, and W_{10} have the same blocking structures. W_4, for example, is given by

$$W_4 = [\underbrace{\xi_1^{(4)} y_4 y_5,\ \xi_1^{(4)} y_5 y_6}_{\mathcal{B}_1},\ \underbrace{\xi_2^{(4)} y_6 y_7,\ \xi_2^{(4)} y_7 y_8,\ \xi_2^{(4)} y_8 y_9}_{\mathcal{B}_2},\ \ldots,\ \underbrace{\xi_{20}^{(4)} y_{60} y_{61},\ \xi_{20}^{(4)} y_{61} y_{62},\ \xi_{20}^{(4)} y_{62} y_{63}}_{\mathcal{B}_{20}}]',$$

where $\xi_1^{(4)}, \ldots, \xi_{20}^{(4)} \overset{i.i.d.}{\sim} N(0,1)$ and they are independent of $\{\xi_1^{(1)}, \ldots, \xi_{20}^{(1)}\}$. Note that $\mathcal{B}_s$ of W_1 and $\mathcal{B}_{s-1}$ of W_4 are scalar multiplications of each other for $s \in \{3, 4, \ldots, 20\}$. Hence bootstrapped autocorrelations in the first and fourth windows follow similar distributions, resulting in similar lower and upper bounds of confidence bands. The same logic applies to W_7 and W_{10}, and hence we observe the periodic behavior in every $b_n = 3$ windows. The same logic follows for an arbitrary block size b_n.

The independence between the ξ's in one window and another is not an essential reason for periodicity at all. By construction, periodicity would emerge even if we were to use the same set of random numbers $\xi_1, \ldots, \xi_{20} \overset{i.i.d.}{\sim} N(0,1)$ for all windows.

Numerical Illustration We now present an illustrative example that supports our argument above. We draw $y_1, y_2, \ldots, y_{71} \overset{i.i.d.}{\sim} N(0,1)$. Set $n = 60$ and $b_n = 3$ as above. We generate 10,000 bootstrap samples in each window.

Figure 4 shows the rolling window sample autocorrelations at lag 1 and their 95% confidence bands computed under $H_0 : \rho(1) = 0$ based on the blockwise wild bootstrap. The confidence bands have a clear periodic pattern with a three-period cycle, which matches block size $b_n = 3$. Given that the data are serially independent, the periodic confidence bands are unfavorable results.

5 Randomizing Block Size

Our solution to the rolling window periodicity of blockwise wild bootstrap confidence bands is to randomize the block size across bootstrap samples and rolling windows. A similar solution to the ingrained non-stationarity of the block bootstrap is proposed by Politis and Romano (1994). See also Lahiri (1999).

In the literature, a conventional choice of block size is $b_n = [c\sqrt{n}]$ with $c \in \{0.5, 1.0, 2.0\}$ (cf. Shao, 2011, Hill and Motegi, 2017). In the main paper Hill and Motegi (2018), we begin with using $c = 1$ and window sizes $n \in \{240, 480, 720\}$ so that $b_n \in \{15, 21, 26\}$ for each block. Then we

independently draw c from the uniform distribution on 0.5 and 1.5 for each bootstrap sample and window, hence $E[c] = 1$. Randomness across windows eliminates the periodicity of confidence bands because any pair of windows no longer has a similar blocking structure. Randomness across bootstrap samples also reduces the variance of confidence bands, which is useful for visual inspection.

More generally, we could draw c from $U(1-\iota, 1+\iota)$. So far we do not have a logical principle on how to choose ι, and that remains as an interesting open question. There is a trade-off that a small ι does not fully eliminate periodicity due to little randomness, while a large ι results in more volatility in confidence bands. In this paper we simply choose $\iota = 0.5$ as a rule of thumb, and verify via controlled experiments in Section 6 and empirical analysis that our choice yields sufficiently non-periodic, smooth confidence bands.

Example 2 *Continue Example 1 and consider the same sample path of* $y_t \overset{i.i.d.}{\sim} N(0,1)$. *We plot sample autocorrelations at lag 1 and their 95% confidence bands based on the blockwise wild bootstrap under* $H_0 : \rho(1) = 0$. *The block size is* $b_n = [c\sqrt{n}]$ *with either* $c = 1$ *(i.e. fixed block size) or* $c \sim U(0.5, 1.5)$ *(i.e. randomized block size). Window sizes are* $n \in \{240, 480, 720\}$. *The number of bootstrap iterations is 5,000 for each window.*

See Figure 5 for results. When we fix $c = 1$, *there is a clear periodic pattern in every* b_n *windows. When we randomize* $c \sim U(0.5, 1.5)$, *the periodicity is eliminated dramatically. This example indicates that randomizing block size is a valid solution to the artificial periodicity in rolling window confidence bands.*

Example 3 *Continue Examples 1 and 2 and consider the same sample path of* $y_t \overset{i.i.d.}{\sim} N(0,1)$. *We perform the Cramér-von Mises test with respect to* $H_0 : \rho(h) = 0$ *with* $h = 1, \ldots, n-1$.[2] *See Figure 6 for test statistics and 95% confidence bands (i.e. 5% critical values) over rolling windows with size* $n \in \{240, 480, 720\}$. *See Figure 7 for associated p-values. Clearly, the fixed block size with* $c = 1$ *results in periodic bootstrapped confidence bands, while the randomized block size with* $c \sim U(0.5, 1.5)$ *results in non-periodic, smooth bands. Compared with the confidence bands, bootstrapped p-values with the fixed block size exhibit much less periodicity. It is still true, however, that block size randomization makes p-values smoother.*

6 Monte Carlo Simulations

In this section, we confirm via Monte Carlo simulations that the blockwise wild bootstrap with randomized block size produces reasonable size and power in finite sample.

[2] Results with the max-correlation and sup-LM tests are qualitatively similar and hence omitted to save space.

6.1 Simulation Design

The simulation design is similar to the one used in Hill and Motegi (2017). Consider a full sample framework with six data generating processes (DGPs).

IID: $y_t = e_t$.	GARCH(1,1): $y_t = \sigma_t e_t,\ \sigma_t^2 = 1 + 0.2y_{t-1}^2 + 0.5\sigma_{t-1}^2$.
Bilinear: $y_t = 0.5e_{t-1}y_{t-2} + e_t$.	MA(1): $y_t = e_t + 0.2e_{t-1}$.
AR(1): $y_t = 0.2y_{t-1} + e_t$.	Remote MA(12): $y_t = e_t + 0.2e_{t-12}$.

For each case, $e_t \overset{i.i.d.}{\sim} N(0,1)$. The white noise property holds for IID, GARCH, and Bilinear and not for MA(1), AR(1), and Remote MA(12). We investigate empirical size for the former cases and empirical power for the latter cases. Sample size is $n \in \{100, 250, 500\}$.

We perform the $\rho(1)$-based test, Hill and Motegi's (2017) max-correlation test, Andrews and Ploberger's (1996) sup-LM test, and Shao's (2011) Cramér-von Mises test. For the max-correlation test, lag length is $\mathcal{L}_n = \max\{5, [\delta n/\ln(n)]\}$ with $\delta \in \{0.0, 0.2, 0.4, 0.5, 1.0\}$. For the sup-LM test we also cover $\mathcal{L}_n = n - 1$. For the CvM test $\mathcal{L}_n = n - 1$.

We execute the blockwise wild bootstrap with block size $b_n = c\sqrt{n}$ for p-value computation. We consider a fixed block size and a randomized block size. For the former, we set $c = 1$ for all $M = 1000$ bootstrap samples. For the latter, we draw $c \sim U(0.5, 1.5)$ independently across $M = 1000$ bootstrap samples.

We report rejection frequencies at the 1%, 5%, and 10% levels after $J = 1000$ Monte Carlo trials.

6.2 Results

See Tables 1-6 for rejection frequencies. Fixed block size and randomized block size produce almost identical rejection frequencies for all six DGPs. Both approaches control for size well (Tables 1-3), and produce reasonable power (Tables 4-6). Those results suggest that the asymptotic validity of the blockwise wild bootstrap is preserved under the randomization of block size. An intuition is that we keep the same order $b_n = c\sqrt{n} = O(n^{1/2})$ no matter what value c takes, and hence the asymptotic validity is preserved.

Under MA(1) and AR(1), the $\rho(1)$-based test is often most powerful (Tables 4-5). This is not surprising since $\rho(h)$ takes a nonzero value *if and only if* $h = 1$ for MA(1), and $\rho(h)$ decays geometrically for AR(1). Since the AP and CvM tests put the most weight on short lags regardless of $\mathcal{L}_n$, their empirical power is close to the $\rho(1)$-based test. The max-correlation test with a large lag length has lower power than the other tests, since it treats short and distant lags equally.

Under remote MA(12), the max-correlation test is by far the most powerful (Table 6). Since $\rho(h)$ takes a nonzero value *if and only if* $h = 12$, the $\rho(1)$-based test naturally has no power beyond size. The same goes for the AP and CvM tests since they put small weights on remote lags. The max-correlation test with a sufficiently large lag length $\mathcal{L}_n \geq 12$ achieves remarkably high power. All of these implications are consistent with simulation results reported in Hill and Motegi (2017).

7 Omitted Empirical Results

In this section we present extra empirical results on stock returns. We first report summary statistics of our stock return data in Table 7. Each log return series of Shanghai, Nikkei, FTSE, and SP500 has a positive mean, but it is not significant at the 5% level according to a bootstrapped confidence band. Shanghai returns have the largest standard deviation of 0.017, but Nikkei returns have the greatest range of $[-0.121, 0.132]$: it has the largest minimum and maximum in absolute value. Each series displays negative skewness and large kurtosis, all stylized traits. Due to the negative skewness and excess kurtosis, the p-values of the Kolmogorov-Smirnov and Anderson-Darling tests of normality are well below 1% for all series, strong evidence against normality.

In Section 7.1 we perform full sample analysis of stock returns, which is entirely omitted in the main paper. In Section 7.2 we report some omitted results on the rolling window analysis.

7.1 Full Sample Analysis

Figure 8 shows sample autocorrelations of the daily return series from January 1, 2003 through October 29, 2015. Lags of $h = 1, \ldots, 25$ trading days are considered. The 95% confidence bands are constructed with Shao's (2011) blockwise wild bootstrap under the null hypothesis of white noise. Hence a sample autocorrelation lying outside the confidence band can be thought of as an evidence *against* the white noise hypothesis, conditional on each lag h. The number of bootstrap samples is $M = 5000$, and in all cases we draw from the standard normal distribution.

While most Shanghai sample correlations lie inside the 95% bands, there are some marginal cases. Correlations at lags 3 and 4, for example, are respectively 0.036 and 0.064, which are on or near the upper bounds. We need a formal white noise test that considers all lags jointly to judge whether the Chinese stock market is weak form efficient.

Nikkei, FTSE, and SP500 have significantly negative correlations at lag 1. The sample correlation with accompanied confidence band is -0.036 with [-0.031, 0.031] for Nikkei; -0.054 with [-0.042, 0.042] for FTSE; -0.102 with [-0.079, 0.075] for SP500. Thus, as far as $h = 1$ is concerned, the Nikkei, FTSE, and SP500 returns are unlikely to be white noise, evidence *against* the weak form efficiency of the Japanese, U.K., and U.S. stock markets.

Table 8 compiles bootstrapped p-values from the max-correlation, sup-LM, and CvM white

noise tests. The max-correlation test requires the maximum lag length $\mathcal{L}_n = o(n)$, while the sup-LM and CvM tests use $\mathcal{L}_n = n - 1$. Nankervis and Savin (2010) truncate the correlation series in the sup-LM statistic, using $\mathcal{L}_n = 20$ for each n, which fails to deliver a consistent test. We use each $\mathcal{L}_n = \max\{5, [\delta \times n/\ln(n)]\}$ with $\delta \in \{0.0, 0.2, 0.4, 0.5, 1.0\}$ for the max-correlation and sup-LM tests for comparability, as well as $\mathcal{L}_n = n - 1$ for sup-LM and CvM tests. Note that, under suitable regularity conditions, as long as $\mathcal{L}_n \to \infty$ then the sup-LM and CvM tests will have their intended limit properties under the null and alternative hypotheses, even if $\mathcal{L}_n = o(n)$. In the case of Shanghai, for example, we have $n = 3110$ and hence $\mathcal{L}_n \in \{5, 77, 154, 193, 386\}$. The bootstrap block size is set to be $b_n = [\sqrt{n}]$ as in Hill and Motegi (2017), but values like $b_n = [0.5\sqrt{n}]$ or $b_n = [2\sqrt{n}]$ lead to similar results (cf. Shao, 2011).

Focusing on the 5% significance level, no tests reject the white noise hypothesis for Shanghai. The smallest p-value for Shanghai arises at the max-correlation test at maximum lag 5, but it is $p = 0.053$. As more lags are added to the max-correlation and sup-LM statistics, the bootstrapped p-values are progressively larger. The CvM test with all possible lags likewise fails to reject the null hypothesis. Since we do not observe any strong evidence against the white noise hypothesis, we conclude that the Chinese stock market is weak form efficient.

Similarly, no tests reject the white noise hypothesis for Nikkei and FTSE at the 5% level. For Nikkei, the smallest p-value of 0.060 is produced by the CvM test. For FTSE, the smallest p-value of 0.065 is produced by the sup-LM test with $\mathcal{L}_n = 5$. We thus conclude that the Japanese and U.K. stock markets are weak form efficient.

For SP500, the white noise hypothesis is rejected at the 5% level by the sup-LM tests with all lags $\mathcal{L}_n \in \{5, 79, 159, 199, 399, 3229\}$ and the CvM test. In view of Figure 8, the rejection stems from the large negative autocorrelation at lag 1. The max-correlation test leads to non-rejections, although we get $p = 0.094$ at $\mathcal{L}_n = 5$. This result is not surprising since the max-correlation test treats each lag equally while the sup-LM and CvM tests put most weights on small lags (cf. Hill and Motegi, 2017). Hence, we conclude that the U.S. stock market is inefficient due to the large negative autocorrelation at lag 1.

7.2 Rolling Window Analysis

Our main interest lies in rolling window analysis of stock returns so that we can capture a potentially time-varying degree of market efficiency. In the main paper, we consider window sizes $n \in \{240, 480, 720\}$ days and perform the CvM test. Our empirical results are summarized as follows. First, the white noise hypothesis is accepted in most windows for Shanghai and Nikkei, indicating that the Chinese and Japanese stock markets are weak form efficient. Second, the

white noise hypothesis is accepted for FTSE and SP500 in non-crisis periods, but often rejected in crisis periods like Iraq War and the subprime mortgage crisis. The reason for those rejections is large negative autocorrelations at lag 1. The U.K. and U.S. stock markets are therefore inefficient during crisis.

In this supplemental material, we add the max-correlation test and the sup-LM test with lags $\mathcal{L}_n = \max\{5, [\delta \times n/\ln(n)]\}$, where $\delta \in \{0.0, 0.2, 0.4, 0.5, 1.0\}$. The specific values of lag length are $\mathcal{L}_n \in \{5, 8, 17, 21, 43\}$ for $n = 240$; $\mathcal{L}_n \in \{5, 15, 31, 38, 77\}$ for $n = 480$; $\mathcal{L}_n \in \{5, 21, 43, 54, 109\}$ for $n = 720$. In addition, we use $\mathcal{L}_n = n - 1$ for the sup-LM test. As in the main paper, block size parameter c is randomly drawn from the uniform distribution $U(0.5, 1.5)$. See Section 2 for the construction of the max-correlation and sup-LM tests.

See Table 9 for summary results. The sup-LM test with any lag length yields a similar result with the CvM test. When $n = 480$, for instance, the ratio of rolling windows where a rejection happens is nearly zero for Shanghai and Nikkei; roughly 0.15 for FTSE; roughly 0.23 for SP500.[3] It is not surprising that the sup-LM and CvM tests produce similar results since both of them put most weights on short lags by construction.

In view of Table 9, the max-correlation test rarely rejects the white noise hypothesis for any series, suggesting that each market is efficient. The max-correlation test is designed to be robust for remote autocorrelations. In our present case, the first lag seems most important and hence it is not surprising that the sup-LM and CvM tests lead to more rejections than the max-correlation test. Overall, we confirm the conclusion of the main paper: the Chinese and Japanese markets are efficient throughout the sample period, while the U.K. and U.S. markets are inefficient in crisis periods due to large negative autocorrelations.

[3] P-values of the max-correlation and sup-LM tests across rolling windows are omitted for brevity, but available upon request.

Table 1: Rejection Frequencies – IID $y_t = e_t$

	$n = 100$		$n = 250$		$n = 500$	
	Fixed	Randomized	Fixed	Randomized	Fixed	Randomized
	1%, 5%, 10%	1%, 5%, 10%	1%, 5%, 10%	1%, 5%, 10%	1%, 5%, 10%	1%, 5%, 10%
$\rho(1)$	.041, .105, .168	.018, .081, .153	.022, .069, .116	.017, .063, .116	.018, .072, .136	.020, .062, .125
Max ($\delta = 0.0$)	.013, .055, .122	.012, .049, .113	.011, .068, .134	.009, .054, .127	.007, .052, .121	.005, .041, .120
Max ($\delta = 0.2$)	.015, .064, .134	.010, .053, .119	.005, .040, .098	.006, .032, .096	.008, .041, .088	.003, .035, .082
Max ($\delta = 0.4$)	.005, .052, .115	.004, .036, .101	.005, .036, .077	.004, .030, .070	.005, .041, .080	.003, .034, .066
Max ($\delta = 0.5$)	.007, .041, .091	.006, .032, .086	.005, .036, .093	.002, .032, .078	.002, .039, .093	.002, .036, .086
Max ($\delta = 1.0$)	.003, .028, .072	.002, .023, .072	.004, .032, .088	.002, .032, .085	.007, .030, .070	.005, .028, .070
AP ($\delta = 0.0$)	.023, .079, .143	.005, .055, .128	.016, .059, .117	.013, .050, .102	.009, .066, .119	.008, .056, .116
AP ($\delta = 0.2$)	.020, .078, .147	.007, .061, .131	.009, .049, .119	.007, .045, .109	.008, .040, .101	.006, .034, .093
AP ($\delta = 0.4$)	.015, .076, .146	.005, .048, .129	.014, .065, .124	.005, .051, .111	.010, .051, .101	.006, .046, .103
AP ($\delta = 0.5$)	.007, .052, .124	.007, .040, .101	.018, .066, .126	.009, .055, .119	.012, .056, .113	.009, .054, .102
AP ($\delta = 1.0$)	.017, .062, .125	.008, .044, .111	.011, .070, .133	.004, .054, .114	.012, .061, .115	.011, .055, .117
AP ($n - 1$)	.016, .055, .127	.003, .036, .094	.010, .049, .117	.006, .042, .104	.009, .071, .129	.006, .058, .123
CvM ($n - 1$)	.018, .081, .136	.009, .069, .131	.020, .063, .113	.015, .054, .108	.010, .049, .105	.010, .049, .103

Rejection frequencies (1%, 5%, 10%) after 1000 Monte Carlo trials. The $\rho(1)$-based test, Hill and Motegi's (2017) max-correlation test, Andrews and Ploberger's (1996) sup-LM test, and Shao's (2011) Cramér-von Mises test are performed. For the max-correlation test, lag length is $\mathcal{L}_n = \max\{5, [\delta n/\ln(n)]\}$ with $\delta \in \{0.0, 0.2, 0.4, 0.5, 1.0\}$. For the AP test we cover the same values of δ and additionally $\mathcal{L}_n = n - 1$. For the CvM test we cover only $\mathcal{L}_n = n - 1$. Block size of the blockwise wild bootstrap is $b_n = c\sqrt{n}$. "Fixed" means $c = 1$, and "Randomized" means that we draw $c \sim U(0.5, 1.5)$ independently across bootstrap samples.

Table 2: Rejection Frequencies – GARCH(1,1) $y_t = \sigma_t e_t$, $\sigma_t^2 = 1.0 + 0.2y_{t-1}^2 + 0.5\sigma_{t-1}^2$

	$n = 100$		$n = 250$		$n = 500$	
	Fixed	Randomized	Fixed	Randomized	Fixed	Randomized
	1%, 5%, 10%	1%, 5%, 10%	1%, 5%, 10%	1%, 5%, 10%	1%, 5%, 10%	1%, 5%, 10%
$\rho(1)$	.021, .085, .134	.005, .052, .108	.013, .062, .112	.006, .046, .112	.018, .076, .125	.012, .063, .121
Max ($\delta = 0.0$)	.005, .057, .130	.004, .036, .122	.007, .048, .105	.004, .030, .099	.013, .060, .116	.010, .058, .112
Max ($\delta = 0.2$)	.005, .050, .117	.003, .028, .102	.009, .053, .113	.002, .045, .094	.003, .038, .089	.003, .026, .085
Max ($\delta = 0.4$)	.005, .041, .110	.001, .032, .104	.001, .032, .082	.001, .030, .080	.003, .019, .071	.004, .018, .064
Max ($\delta = 0.5$)	.004, .039, .102	.001, .031, .085	.004, .023, .077	.002, .025, .063	.001, .024, .072	.000, .025, .069
Max ($\delta = 1.0$)	.002, .024, .089	.001, .025, .078	.001, .028, .076	.000, .023, .069	.002, .023, .059	.001, .018, .057
AP ($\delta = 0.0$)	.019, .068, .131	.006, .039, .108	.008, .046, .116	.003, .039, .097	.017, .056, .116	.014, .050, .112
AP ($\delta = 0.2$)	.012, .054, .129	.007, .043, .105	.012, .051, .118	.005, .040, .097	.010, .066, .128	.005, .051, .117
AP ($\delta = 0.4$)	.020, .079, .148	.004, .053, .116	.014, .052, .103	.004, .041, .096	.013, .057, .104	.008, .043, .091
AP ($\delta = 0.5$)	.016, .057, .114	.005, .037, .098	.005, .056, .111	.006, .045, .101	.009, .064, .121	.006, .054, .112
AP ($\delta = 1.0$)	.009, .056, .123	.006, .030, .101	.013, .059, .114	.009, .046, .111	.009, .053, .096	.004, .047, .103
AP ($n - 1$)	.010, .055, .122	.000, .031, .095	.013, .061, .110	.010, .047, .109	.007, .046, .093	.006, .029, .079
CvM ($n - 1$)	.013, .066, .137	.008, .054, .128	.017, .062, .122	.009, .055, .106	.013, .053, .120	.009, .061, .116

Rejection frequencies (1%, 5%, 10%) after 1000 Monte Carlo trials. The $\rho(1)$-based test, Hill and Motegi's (2017) max-correlation test, Andrews and Ploberger's (1996) sup-LM test, and Shao's (2011) Cramér-von Mises test are performed. For the max-correlation test, lag length is $\mathcal{L}_n = \max\{5, [\delta n/\ln(n)]\}$ with $\delta \in \{0.0, 0.2, 0.4, 0.5, 1.0\}$. For the AP test we cover the same values of δ and additionally $\mathcal{L}_n = n - 1$. For the CvM test we cover only $\mathcal{L}_n = n - 1$. Block size of the blockwise wild bootstrap is $b_n = c\sqrt{n}$. "Fixed" means $c = 1$, and "Randomized" means that we draw $c \sim U(0.5, 1.5)$ independently across bootstrap samples.

Table 3: Rejection Frequencies – Bilinear $y_t = 0.50e_{t-1}y_{t-2} + e_t$

	$n = 100$		$n = 250$		$n = 500$	
	Fixed	Randomized	Fixed	Randomized	Fixed	Randomized
	1%, 5%, 10%	1%, 5%, 10%	1%, 5%, 10%	1%, 5%, 10%	1%, 5%, 10%	1%, 5%, 10%
$\rho(1)$	.028, .106, .180	.025, .078, .157	.027, .092, .148	.025, .079, .144	.017, .069, .115	.012, .068, .109
Max ($\delta = 0.0$)	.009, .063, .139	.004, .048, .137	.010, .061, .139	.004, .055, .133	.011, .062, .119	.007, .043, .113
Max ($\delta = 0.2$)	.016, .065, .125	.009, .049, .125	.004, .040, .104	.004, .025, .090	.005, .035, .080	.005, .035, .072
Max ($\delta = 0.4$)	.007, .047, .115	.002, .034, .105	.004, .031, .080	.002, .029, .075	.004, .033, .095	.004, .034, .097
Max ($\delta = 0.5$)	.006, .044, .109	.003, .032, .096	.004, .033, .086	.003, .027, .080	.004, .027, .074	.002, .027, .072
Max ($\delta = 1.0$)	.002, .034, .096	.002, .024, .091	.001, .022, .077	.002, .017, .064	.004, .022, .057	.004, .021, .055
AP ($\delta = 0.0$)	.023, .076, .158	.010, .057, .141	.019, .077, .136	.014, .079, .133	.019, .064, .127	.010, .058, .112
AP ($\delta = 0.2$)	.015, .058, .129	.004, .046, .118	.016, .066, .131	.006, .063, .128	.020, .052, .120	.010, .048, .105
AP ($\delta = 0.4$)	.018, .078, .155	.003, .052, .144	.020, .068, .128	.014, .055, .112	.011, .041, .115	.013, .043, .107
AP ($\delta = 0.5$)	.022, .086, .161	.009, .071, .146	.016, .071, .132	.008, .053, .122	.014, .063, .121	.011, .049, .116
AP ($\delta = 1.0$)	.024, .083, .155	.012, .060, .129	.018, .067, .120	.013, .052, .106	.012, .061, .112	.011, .043, .102
AP ($n - 1$)	.015, .082, .156	.006, .052, .133	.014, .066, .123	.006, .050, .110	.010, .054, .117	.008, .046, .107
CvM ($n - 1$)	.028, .106, .168	.014, .075, .161	.024, .085, .142	.019, .077, .136	.022, .072, .126	.015, .059, .115

Rejection frequencies (1%, 5%, 10%) after 1000 Monte Carlo trials. The $\rho(1)$-based test, Hill and Motegi's (2017) max-correlation test, Andrews and Ploberger's (1996) sup-LM test, and Shao's (2011) Cramér-von Mises test are performed. For the max-correlation test, lag length is $\mathcal{L}_n = \max\{5, [\delta n/\ln(n)]\}$ with $\delta \in \{0.0, 0.2, 0.4, 0.5, 1.0\}$. For the AP test we cover the same values of δ and additionally $\mathcal{L}_n = n - 1$. For the CvM test we cover only $\mathcal{L}_n = n - 1$. Block size of the blockwise wild bootstrap is $b_n = c\sqrt{n}$. "Fixed" means $c = 1$, and "Randomized" means that we draw $c \sim U(0.5, 1.5)$ independently across bootstrap samples.

Table 4: Rejection Frequencies – MA(1) $y_t = e_t + 0.20e_{t-1}$

	$n = 100$		$n = 250$		$n = 500$	
	Fixed	Randomized	Fixed	Randomized	Fixed	Randomized
	1%, 5%, 10%	1%, 5%, 10%	1%, 5%, 10%	1%, 5%, 10%	1%, 5%, 10%	1%, 5%, 10%
$\rho(1)$	.268, .461, .597	.216, .447, .592	.670, .860, .925	.646, .869, .921	.956, .987, .993	.950, .988, .994
Max ($\delta = 0.0$)	.093, .254, .395	.066, .231, .376	.397, .673, .788	.395, .664, .795	.859, .959, .982	.861, .956, .983
Max ($\delta = 0.2$)	.071, .238, .387	.056, .221, .366	.303, .587, .711	.282, .578, .728	.744, .898, .943	.733, .903, .951
Max ($\delta = 0.4$)	.054, .180, .299	.031, .161, .289	.208, .467, .628	.194, .481, .618	.653, .838, .897	.651, .839, .903
Max ($\delta = 0.5$)	.039, .167, .321	.032, .158, .301	.158, .420, .580	.154, .420, .590	.631, .834, .898	.631, .835, .900
Max ($\delta = 1.0$)	.018, .097, .226	.010, .100, .225	.122, .344, .495	.114, .333, .503	.582, .803, .891	.576, .800, .886
AP ($\delta = 0.0$)	.160, .387, .541	.114, .373, .529	.486, .743, .857	.436, .732, .862	.880, .977, .994	.886, .979, .993
AP ($\delta = 0.2$)	.139, .360, .488	.095, .326, .493	.441, .729, .854	.405, .734, .850	.856, .971, .988	.837, .973, .993
AP ($\delta = 0.4$)	.123, .338, .504	.076, .297, .487	.403, .710, .831	.364, .695, .831	.823, .965, .987	.829, .966, .988
AP ($\delta = 0.5$)	.159, .324, .485	.089, .308, .493	.437, .726, .840	.405, .715, .846	.840, .956, .985	.846, .964, .981
AP ($\delta = 1.0$)	.139, .354, .524	.096, .313, .502	.420, .732, .862	.370, .726, .864	.845, .963, .989	.843, .968, .990
AP ($n - 1$)	.133, .321, .470	.095, .293, .465	.433, .728, .853	.403, .727, .854	.848, .973, .994	.832, .976, .994
CvM ($n - 1$)	.201, .422, .535	.158, .398, .545	.637, .857, .922	.604, .849, .923	.949, .989, .996	.947, .989, .996

Rejection frequencies (1%, 5%, 10%) after 1000 Monte Carlo trials. The $\rho(1)$-based test, Hill and Motegi's (2017) max-correlation test, Andrews and Ploberger's (1996) sup-LM test, and Shao's (2011) Cramér-von Mises test are performed. For the max-correlation test, lag length is $\mathcal{L}_n = \max\{5, [\delta n/\ln(n)]\}$ with $\delta \in \{0.0, 0.2, 0.4, 0.5, 1.0\}$. For the AP test we cover the same values of δ and additionally $\mathcal{L}_n = n - 1$. For the CvM test we cover only $\mathcal{L}_n = n - 1$. Block size of the blockwise wild bootstrap is $b_n = c\sqrt{n}$. "Fixed" means $c = 1$, and "Randomized" means that we draw $c \sim U(0.5, 1.5)$ independently across bootstrap samples.

Table 5: Rejection Frequencies – AR(1) $y_t = 0.20y_{t-1} + e_t$

	$n = 100$		$n = 250$		$n = 500$	
	Fixed	Randomized	Fixed	Randomized	Fixed	Randomized
	1%, 5%, 10%	1%, 5%, 10%	1%, 5%, 10%	1%, 5%, 10%	1%, 5%, 10%	1%, 5%, 10%
$\rho(1)$	.267, .488, .598	.215, .465, .605	.655, .845, .914	.640, .843, .912	.960, .992, .998	.967, .991, .998
Max ($\delta = 0.0$)	.080, .228, .392	.057, .221, .381	.445, .684, .801	.398, .686, .799	.873, .957, .977	.860, .961, .979
Max ($\delta = 0.2$)	.100, .274, .418	.075, .265, .407	.352, .605, .744	.318, .611, .736	.764, .915, .958	.764, .913, .957
Max ($\delta = 0.4$)	.063, .212, .346	.056, .199, .344	.238, .500, .649	.231, .495, .638	.690, .877, .929	.692, .873, .933
Max ($\delta = 0.5$)	.041, .192, .340	.029, .172, .332	.230, .495, .629	.214, .483, .641	.693, .851, .902	.683, .853, .906
Max ($\delta = 1.0$)	.022, .109, .214	.020, .090, .214	.172, .392, .526	.154, .389, .523	.627, .817, .873	.610, .823, .870
AP ($\delta = 0.0$)	.155, .370, .507	.111, .336, .520	.499, .759, .874	.466, .762, .877	.889, .984, .994	.891, .986, .997
AP ($\delta = 0.2$)	.156, .346, .488	.105, .319, .494	.434, .741, .849	.388, .739, .860	.840, .966, .981	.836, .959, .985
AP ($\delta = 0.4$)	.114, .336, .504	.078, .322, .478	.432, .712, .849	.403, .722, .858	.835, .966, .985	.839, .962, .991
AP ($\delta = 0.5$)	.130, .327, .476	.072, .275, .473	.414, .716, .841	.385, .720, .833	.854, .968, .983	.856, .972, .985
AP ($\delta = 1.0$)	.128, .335, .482	.077, .306, .486	.449, .730, .845	.408, .732, .850	.840, .965, .988	.833, .973, .989
AP ($n - 1$)	.119, .315, .482	.071, .298, .476	.418, .724, .847	.358, .723, .848	.827, .967, .991	.836, .974, .991
CvM ($n - 1$)	.247, .451, .596	.193, .455, .593	.615, .819, .904	.579, .815, .900	.938, .993, .997	.934, .991, .997

Rejection frequencies (1%, 5%, 10%) after 1000 Monte Carlo trials. The $\rho(1)$-based test, Hill and Motegi's (2017) max-correlation test, Andrews and Ploberger's (1996) sup-LM test, and Shao's (2011) Cramér-von Mises test are performed. For the max-correlation test, lag length is $\mathcal{L}_n = \max\{5, [\delta n/\ln(n)]\}$ with $\delta \in \{0.0, 0.2, 0.4, 0.5, 1.0\}$. For the AP test we cover the same values of δ and additionally $\mathcal{L}_n = n - 1$. For the CvM test we cover only $\mathcal{L}_n = n - 1$. Block size of the blockwise wild bootstrap is $b_n = c\sqrt{n}$. "Fixed" means $c = 1$, and "Randomized" means that we draw $c \sim U(0.5, 1.5)$ independently across bootstrap samples.

Table 6: Rejection Frequencies – Remote MA(12) $y_t = e_t + 0.20e_{t-12}$

	$n = 100$		$n = 250$		$n = 500$	
	Fixed	Randomized	Fixed	Randomized	Fixed	Randomized
	1%, 5%, 10%	1%, 5%, 10%	1%, 5%, 10%	1%, 5%, 10%	1%, 5%, 10%	1%, 5%, 10%
$\rho(1)$	.028, .100, .170	.017, .083, .150	.029, .075, .143	.017, .064, .135	.022, .071, .122	.018, .069, .116
Max ($\delta = 0.0$)	.010, .067, .135	.004, .056, .127	.015, .063, .136	.013, .065, .131	.015, .079, .155	.008, .069, .146
Max ($\delta = 0.2$)	.016, .083, .155	.009, .060, .143	.009, .078, .151	.008, .071, .150	.702, .879, .944	.690, .880, .941
Max ($\delta = 0.4$)	.015, .073, .143	.013, .065, .137	.208, .427, .573	.180, .433, .563	.606, .823, .886	.610, .824, .889
Max ($\delta = 0.5$)	.008, .064, .136	.005, .052, .126	.193, .430, .565	.169, .420, .560	.583, .804, .879	.567, .800, .883
Max ($\delta = 1.0$)	.018, .108, .218	.016, .099, .220	.147, .347, .491	.135, .342, .487	.551, .760, .833	.546, .760, .838
AP ($\delta = 0.0$)	.023, .090, .171	.009, .059, .161	.016, .065, .128	.010, .055, .124	.017, .070, .127	.015, .064, .120
AP ($\delta = 0.2$)	.023, .078, .152	.007, .063, .136	.010, .057, .121	.009, .054, .112	.012, .058, .106	.006, .050, .102
AP ($\delta = 0.4$)	.024, .080, .152	.008, .060, .139	.009, .050, .108	.005, .041, .107	.013, .065, .117	.013, .055, .117
AP ($\delta = 0.5$)	.018, .080, .151	.013, .062, .134	.010, .063, .134	.008, .053, .122	.013, .057, .117	.010, .049, .119
AP ($\delta = 1.0$)	.023, .071, .157	.015, .049, .132	.015, .068, .138	.005, .055, .134	.014, .076, .159	.013, .073, .138
AP ($n - 1$)	.015, .078, .141	.004, .060, .124	.026, .081, .145	.009, .070, .133	.015, .056, .122	.008, .051, .109
CvM ($n - 1$)	.029, .085, .159	.016, .072, .145	.022, .076, .121	.018, .064, .120	.025, .069, .136	.013, .064, .133

Rejection frequencies (1%, 5%, 10%) after 1000 Monte Carlo trials. The $\rho(1)$-based test, Hill and Motegi's (2017) max-correlation test, Andrews and Ploberger's (1996) sup-LM test, and Shao's (2011) Cramér-von Mises test are performed. For the max-correlation test, lag length is $\mathcal{L}_n = \max\{5, [\delta n/\ln(n)]\}$ with $\delta \in \{0.0, 0.2, 0.4, 0.5, 1.0\}$. Given $n = 100$, $\mathcal{L}_n \geq 12$ when $\delta = 1.0$; given $n = 250$, $\mathcal{L}_n \geq 12$ when $\delta \geq 0.4$; given $n = 500$, $\mathcal{L}_n \geq 12$ when $\delta \geq 0.2$. For the AP test we cover the same values of δ and additionally $\mathcal{L}_n = n - 1$. For the CvM test we cover only $\mathcal{L}_n = n - 1$. Block size of the blockwise wild bootstrap is $b_n = c\sqrt{n}$. "Fixed" means $c = 1$, and "Randomized" means that we draw $c \sim U(0.5, 1.5)$ independently across bootstrap samples.

Table 7: Sample Statistics of Log Returns of Stock Price Indices (01/01/2003 - 10/29/2015)

	Shanghai	Nikkei	FTSE	SP500
# Observations	3110	3149	3243	3230
Mean ($\times 10^{-4}$)	2.9	2.5	1.5	2.7
95% Band ($\times 10^{-4}$)	$[-7.4, 7.2]$	$[-5.0, 5.1]$	$[-2.7, 2.6]$	$[-3.8, 3.7]$
Median	0.001	0.001	0.001	0.001
Std. Dev.	0.017	0.015	0.012	0.012
Minimum	-0.093	-0.121	-0.093	-0.095
Maximum	0.090	0.132	0.094	0.110
Skewness	-0.425	-0.532	-0.133	-0.319
Kurtosis	6.831	10.68	11.14	14.02
p-KS	0.000	0.000	0.000	0.000
p-AD	0.001	0.001	0.001	0.001

"95% Band" is a bootstrapped 95% confidence band for the sample mean. It is constructed under the null hypothesis of zero-mean white noise, using the blockwise wild bootstrap with block size $b_n = \sqrt{n}$. The number of bootstrap samples is $M = 10000$. "p-KS" signifies a p-value of the Kolmogorov-Smirnov test, while "p-AD" signifies a p-value of the Anderson-Darling test.

Table 8: P-Values of White Noise Tests in Full Sample (01/01/2003 - 10/29/2015)

	Shanghai		Nikkei		FTSE		SP500	
	$\mathcal{L}_n$	p	$\mathcal{L}_n$	p	$\mathcal{L}_n$	p	$\mathcal{L}_n$	p
Max ($\delta = 0.0$)	5	0.053	5	0.327	5	0.425	5	0.094
Max ($\delta = 0.2$)	77	0.314	78	0.612	80	0.517	79	0.162
Max ($\delta = 0.4$)	154	0.380	156	0.724	160	0.563	159	0.169
Max ($\delta = 0.5$)	193	0.398	195	0.681	200	0.571	199	0.177
Max ($\delta = 1.0$)	386	0.447	390	0.671	401	0.576	399	0.171
AP ($\delta = 0.0$)	5	0.182	5	0.083	5	0.065	5	0.034
AP ($\delta = 0.2$)	77	0.149	78	0.145	80	0.079	79	0.032
AP ($\delta = 0.4$)	154	0.148	156	0.142	160	0.074	159	0.032
AP ($\delta = 0.5$)	193	0.157	195	0.156	200	0.075	199	0.030
AP ($\delta = 1.0$)	386	0.153	390	0.145	401	0.076	399	0.029
AP ($\mathcal{L}_n = n - 1$)	3109	0.161	3148	0.147	3242	0.073	3229	0.028
CvM ($\mathcal{L}_n = n - 1$)	3109	0.154	3148	0.060	3242	0.068	3229	0.034

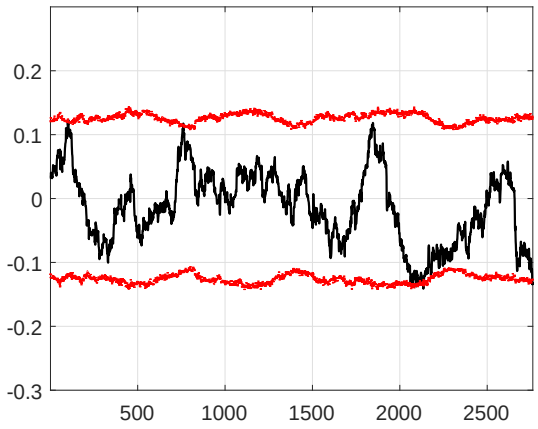
Bootstrapped p-values of Hill and Motegi's (2017) max-correlation white noise test, Andrews and Ploberger's (1996) sup-LM test, and the Cramér-von Mises test. Shao's (2011) blockwise wild bootstrap with 5000 replications is used for each test. The maximum lag lengths for the max-correlation and sup-LM tests are $\mathcal{L}_n = \max\{5, [\delta \times n/\ln(n)]\}$ with $\delta \in \{0.0, 0.2, 0.4, 0.5, 1.0\}$. The sup-LM test is also computed with the maximum possible $n - 1$ lags. The CvM test uses $\mathcal{L}_n = n - 1$.

Table 9: Rejection Ratio of White Noise Tests across Rolling Windows

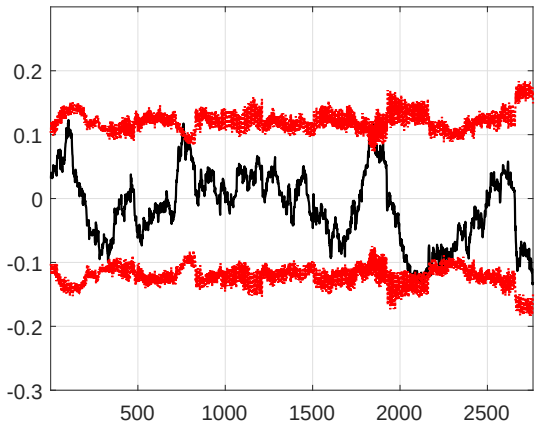
	$n = 240$		$n = 480$		$n = 720$	
	$\mathcal{L}_n$	SH, NK, FT, SP	$\mathcal{L}_n$	SH, NK, FT, SP	$\mathcal{L}_n$	SH, NK, FT, SP
Max ($\delta = 0.0$)	5	.037, .032, .073, .088	5	.010, .007, .087, .040	5	.061, .000, .117, .092
Max ($\delta = 0.2$)	8	.025, .040, .043, .031	15	.009, .000, .017, .011	21	.022, .000, .033, .008
Max ($\delta = 0.4$)	17	.003, .000, .017, .016	31	.006, .000, .009, .008	43	.005, .000, .018, .005
Max ($\delta = 0.5$)	21	.002, .000, .016, .014	38	.004, .000, .000, .005	54	.005, .000, .017, .098
Max ($\delta = 1.0$)	43	.000, .001, .045, .007	77	.002, .000, .018, .032	109	.003, .000, .004, .090
AP ($\delta = 0.0$)	5	.002, .016, .135, .164	5	.001, .005, .154, .226	5	.013, .000, .318, .382
AP ($\delta = 0.2$)	8	.006, .014, .143, .164	15	.002, .002, .146, .236	21	.012, .000, .307, .386
AP ($\delta = 0.4$)	17	.003, .020, .132, .159	31	.002, .002, .149, .237	43	.013, .000, .311, .386
AP ($\delta = 0.5$)	21	.003, .021, .131, .158	38	.001, .003, .150, .236	54	.012, .000, .310, .386
AP ($\delta = 1.0$)	43	.002, .021, .130, .158	77	.002, .003, .150, .236	109	.012, .000, .310, .386
AP ($\mathcal{L}_n = n - 1$)	239	.003, .020, .130, .158	479	.001, .003, .149, .234	719	.011, .000, .307, .387
CvM ($\mathcal{L}_n = n - 1$)	239	.058, .001, .143, .213	479	.010, .000, .171, .266	719	.023, .000, .227, .443

The ratio of rolling windows where the null hypothesis of white noise is rejected at the 5% level. Shanghai (SH), Nikkei (NK), FTSE (FT), and S&P 500 (SP) are analyzed. Window size is $n \in \{240, 480, 720\}$ days. Test statistics are Hill and Motegi's (2017) max-correlation statistic (Max), Andrews and Ploberger's (1996) sup-LM statistic (AP), and the Cramér-von Mises statistic (CvM). Lag length is $\mathcal{L}_n = \min\{5, [\delta \times n / \ln n]\}$ for the Max and AP tests. For the AP test, we also use $\mathcal{L}_n = n - 1$. For the CvM test, we only use $\mathcal{L}_n = n - 1$. We use Shao's (2011) blockwise wild bootstrap with block size $b_n = [c \times \sqrt{n}]$. We draw $c \sim U(0.5, 1.5)$ independently across $M = 5000$ bootstrap samples and rolling windows.

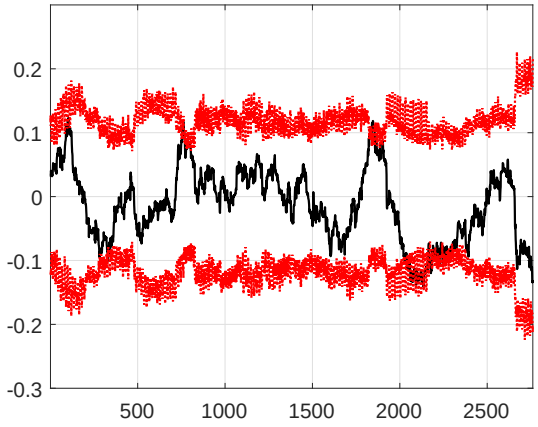
Figure 1: Confidence Bands for Autocorrelations with Window Size $n = 240$



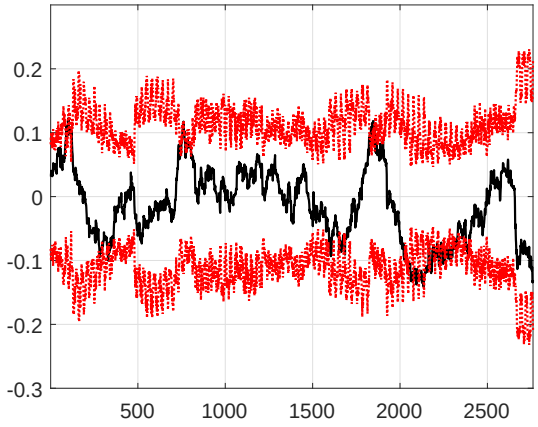
Wild Bootstrap



BWB with $c = 0.5$ ($b_n = 7$)



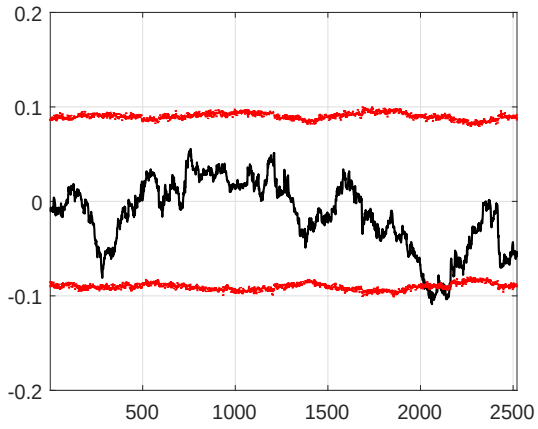
BWB with $c = 1$ ($b_n = 15$)



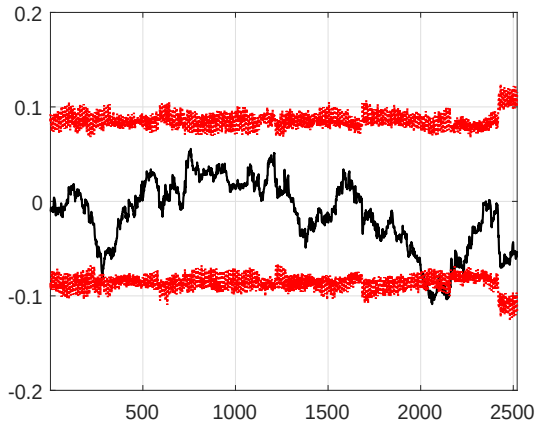
BWB with $c = 2$ ($b_n = 30$)

The data generating process is $y_t \overset{i.i.d.}{\sim} N(0,1)$ with sample size $T = 3000$. This figure plots rolling window sample autocorrelations at lag 1 and their 95% confidence bands based on the wild bootstrap and blockwise wild bootstrap (BWB) under $H_0 : \rho(1) = 0$. For BWB, block size is $b_n = [c\sqrt{n}]$ with $c \in \{0.5, 1, 2\}$ and hence $b_n \in \{7, 15, 30\}$. The number of bootstrap iterations is 5,000 for each window. Each point on the horizontal axis represents the window ID number.

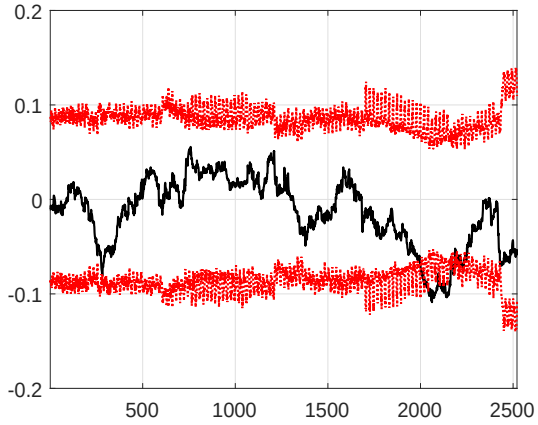
Figure 2: Confidence Bands for Autocorrelations with Window Size $n = 480$



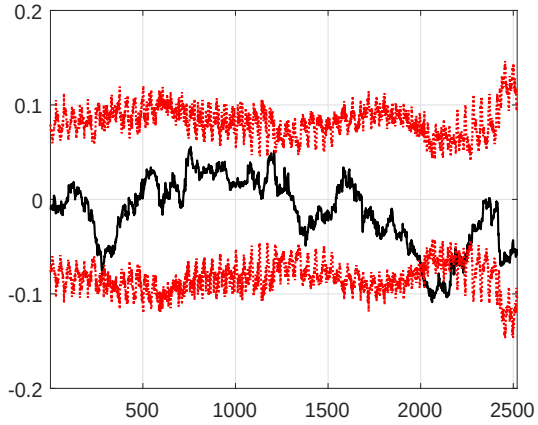
Wild Bootstrap



BWB with $c = 0.5$ ($b_n = 10$)



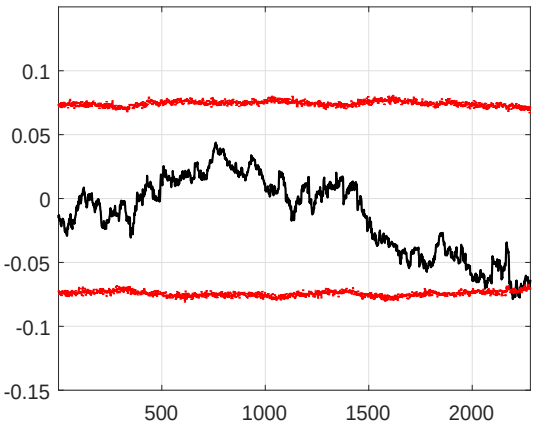
BWB with $c = 1$ ($b_n = 21$)



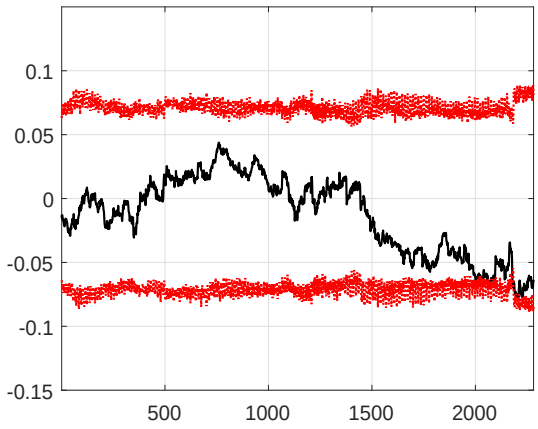
BWB with $c = 2$ ($b_n = 43$)

The data generating process is $y_t \overset{i.i.d.}{\sim} N(0,1)$ with sample size $T = 3000$. This figure plots rolling window sample autocorrelations at lag 1 and their 95% confidence bands based on the wild bootstrap and blockwise wild bootstrap (BWB) under $H_0 : \rho(1) = 0$. For BWB, block size is $b_n = [c\sqrt{n}]$ with $c \in \{0.5, 1, 2\}$ and hence $b_n \in \{10, 21, 43\}$. The number of bootstrap iterations is 5,000 for each window. Each point on the horizontal axis represents the window ID number.

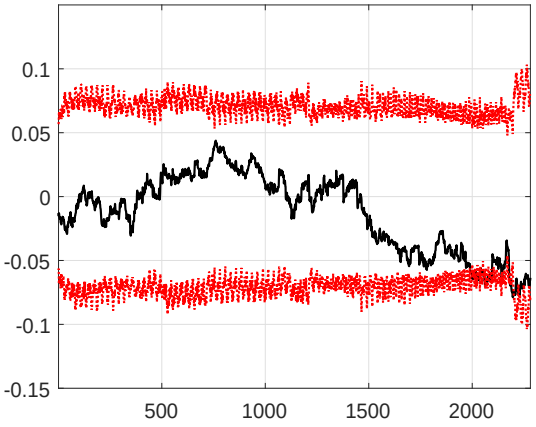
Figure 3: Confidence Bands for Autocorrelations with Window Size $n = 720$



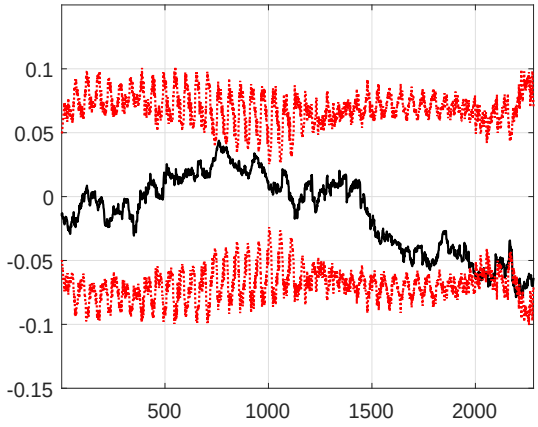
Wild Bootstrap



BWB with $c = 0.5$ ($b_n = 13$)



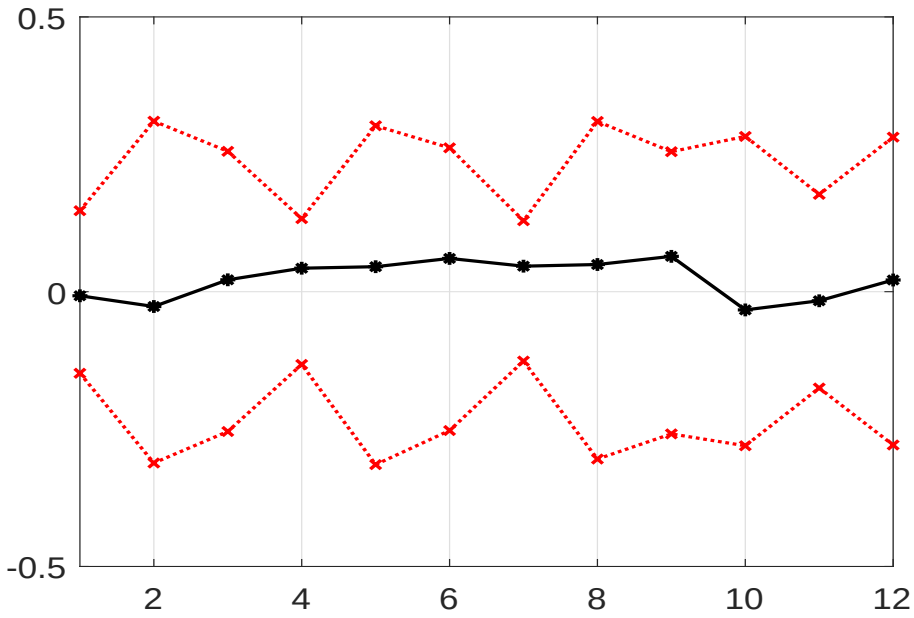
BWB with $c = 1$ ($b_n = 26$)



BWB with $c = 2$ ($b_n = 53$)

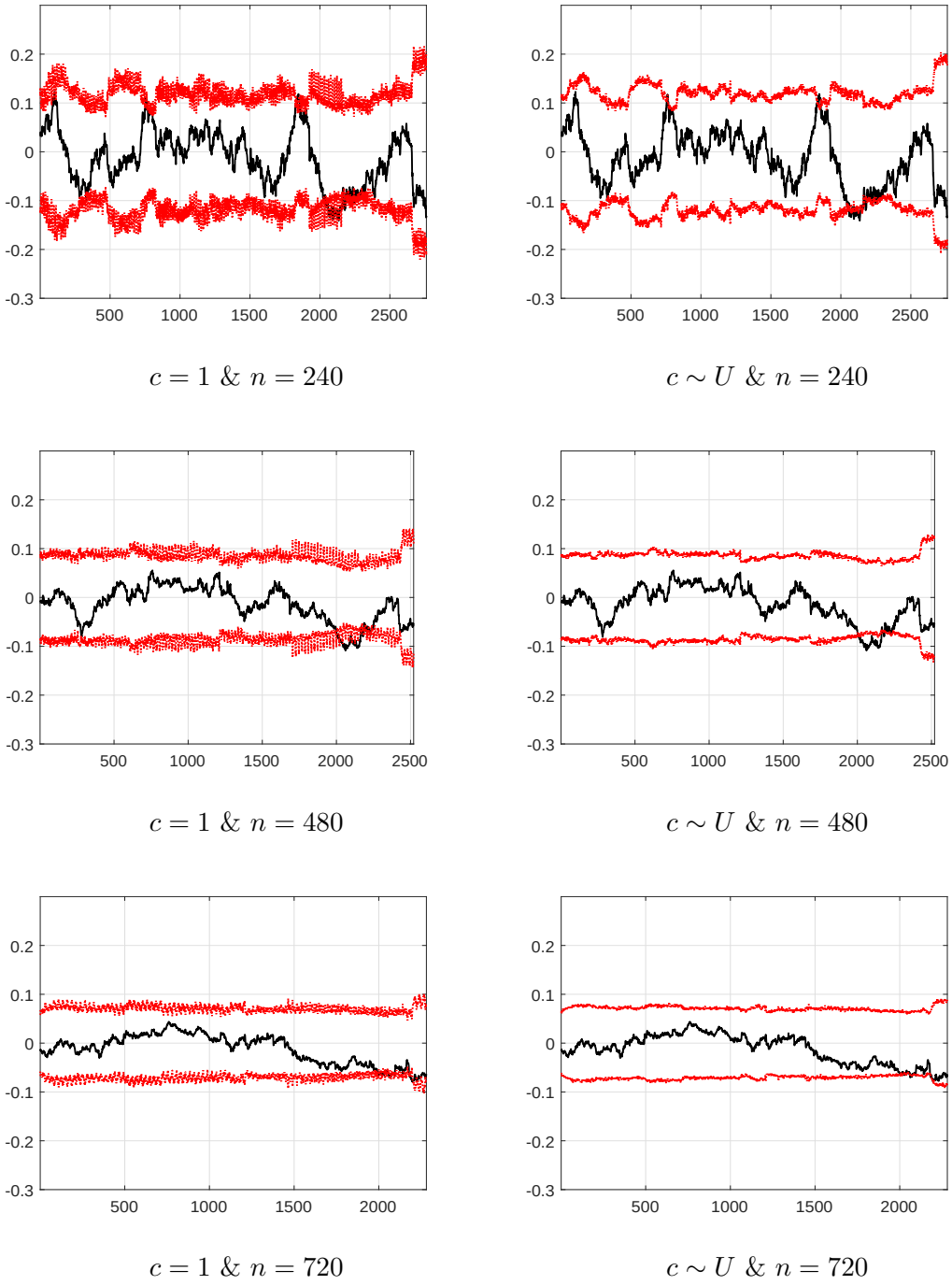
The data generating process is $y_t \overset{i.i.d.}{\sim} N(0,1)$ with sample size $T = 3000$. This figure plots rolling window sample autocorrelations at lag 1 and their 95% confidence bands based on the wild bootstrap and blockwise wild bootstrap (BWB) under $H_0 : \rho(1) = 0$. For BWB, block size is $b_n = [c\sqrt{n}]$ with $c \in \{0.5, 1, 2\}$ and hence $b_n \in \{13, 26, 53\}$. The number of bootstrap iterations is 5,000 for each window. Each point on the horizontal axis represents the window ID number.

Figure 4: Illustrative Example



The data generating process is $y_t \overset{i.i.d.}{\sim} N(0,1)$ with sample size $T = 71$. Window size is $n = 60$, resulting in 12 windows. This figure plots the rolling window sample autocorrelations of $\{y_t\}$ at lag 1 and their 95% confidence bands based on the blockwise wild bootstrap with block size $b_n = 3$. We generate 10,000 bootstrap samples in each window.

Figure 5: Confidence Bands for Autocorrelations with Fixed versus Randomized Block Sizes



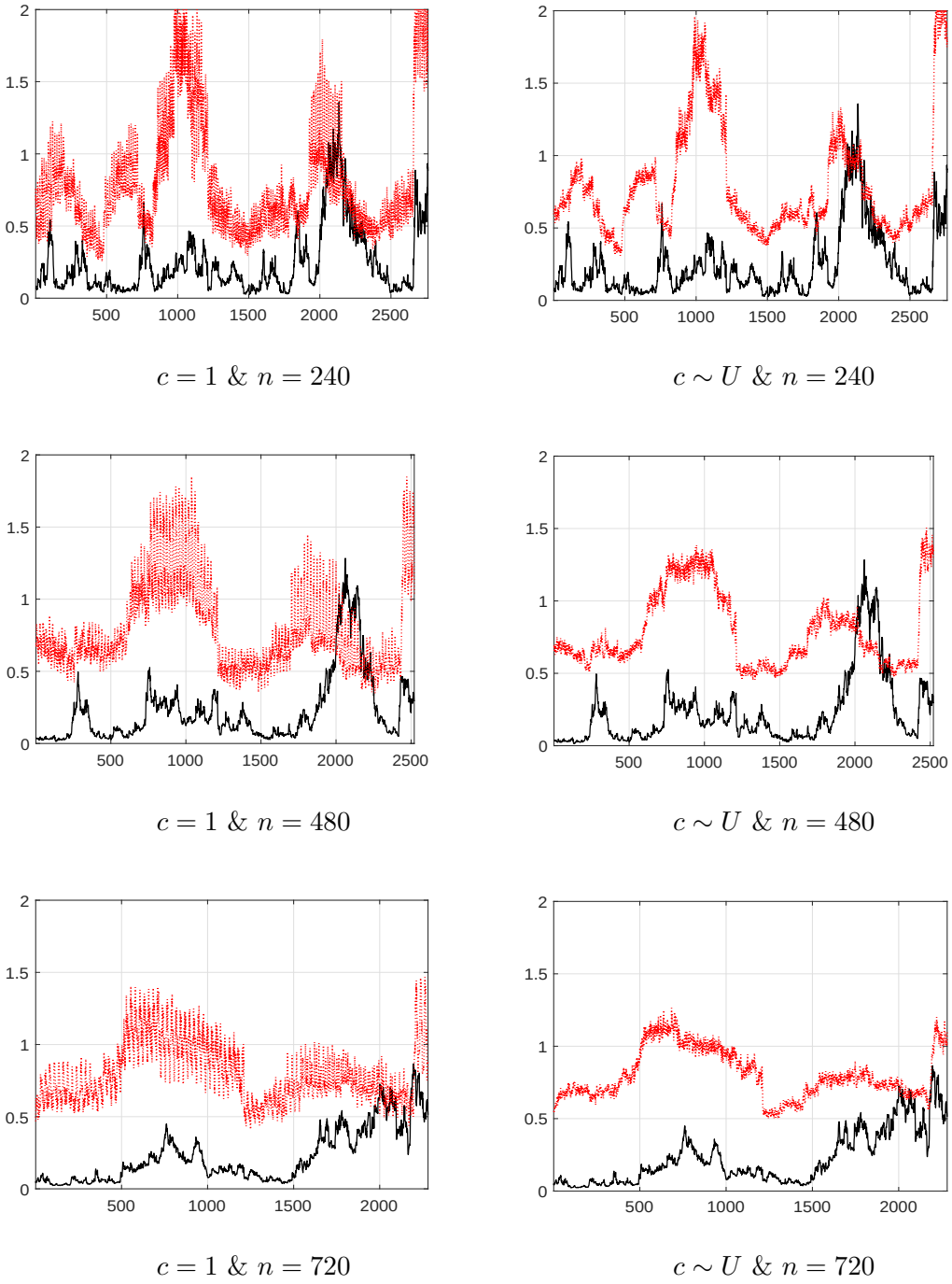
$c = 1$ & $n = 240$ $\quad$ $c \sim U$ & $n = 240$

$c = 1$ & $n = 480$ $\quad$ $c \sim U$ & $n = 480$

$c = 1$ & $n = 720$ $\quad$ $c \sim U$ & $n = 720$

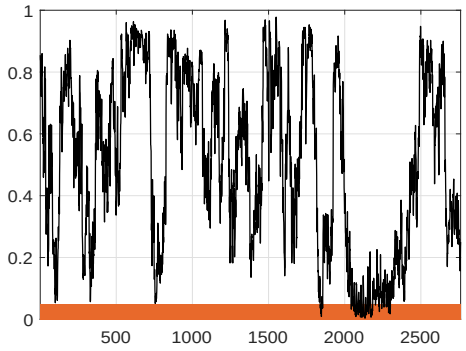
The data generating process is $y_t \overset{i.i.d.}{\sim} N(0,1)$ with sample size $T = 3000$. This figure plots rolling window sample autocorrelations at lag 1 and their 95% confidence bands based on the blockwise wild bootstrap under $H_0 : \rho(1) = 0$. The block size is $b_n = [c\sqrt{n}]$ with either $c = 1$ (i.e. fixed block size) or $c \sim U(0.5, 1.5)$ (i.e. randomized block size). Window size is $n \in \{240, 480, 720\}$. The number of bootstrap iterations is 5,000 for each window. Each point on the horizontal axis represents the window ID number.

Figure 6: Cramér-von Mises Test Statistics with Fixed versus Randomized Block Sizes

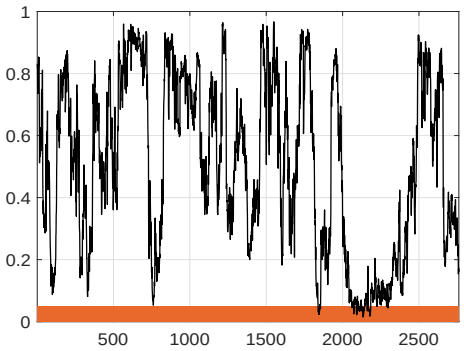


The data generating process is $y_t \overset{i.i.d.}{\sim} N(0,1)$ with sample size $T = 3000$. We perform rolling window Cramér-von Mises tests based on the blockwise wild bootstrap under $H_0 : \rho(h) = 0$ for $h \geq 1$. This figure plots test statistics in black, solid lines and 5% critical values (i.e. 95% confidence bands) in red, dotted lines. The block size is $b_n = [c\sqrt{n}]$ with either $c = 1$ (i.e. fixed block size) or $c \sim U(0.5, 1.5)$ (i.e. randomized block size). Window size is $n \in \{240, 480, 720\}$. The number of bootstrap iterations is 5,000 for each window. Each point on the horizontal axis represents the window ID number.

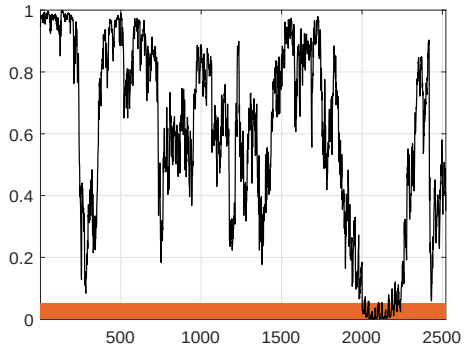
Figure 7: P-Values of Cramér-von Mises Test with Fixed versus Randomized Block Sizes



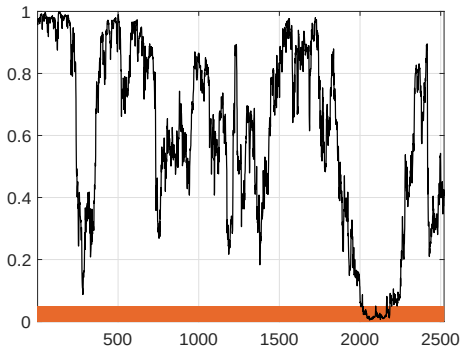
$c = 1$ & $n = 240$



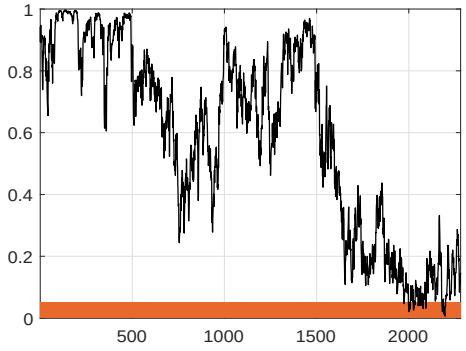
$c \sim U$ & $n = 240$



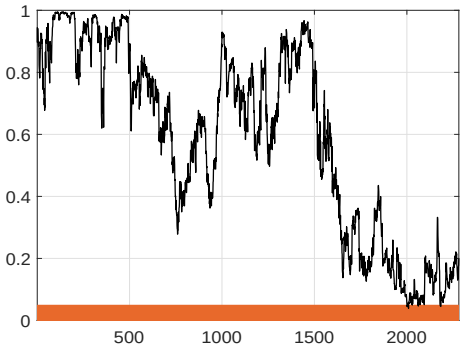
$c = 1$ & $n = 480$



$c \sim U$ & $n = 480$



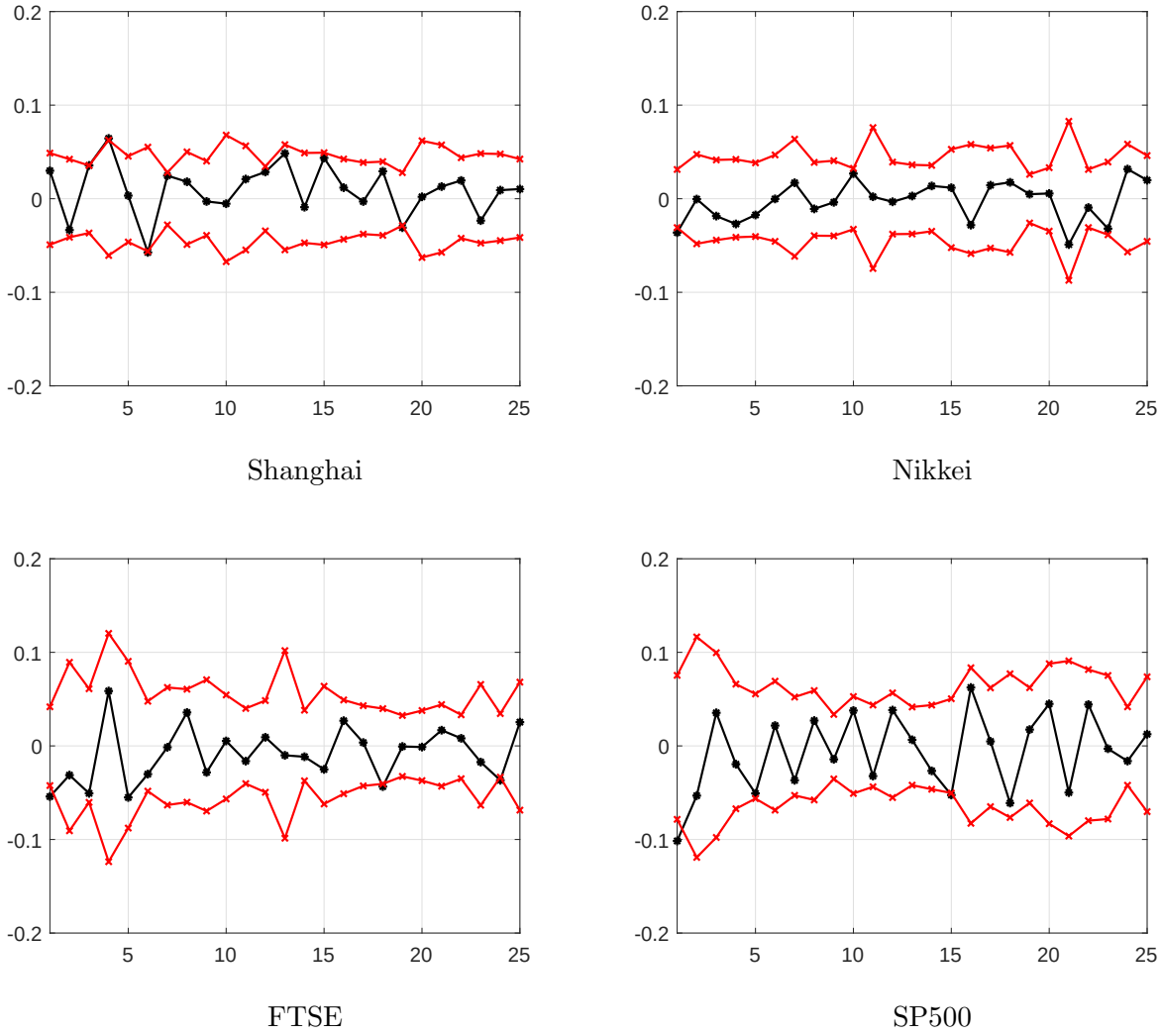
$c = 1$ & $n = 720$



$c \sim U$ & $n = 720$

The data generating process is $y_t \overset{i.i.d.}{\sim} N(0,1)$ with sample size $T = 3000$. This figure plots rolling window p-values of Cramér-von Mises tests based on the blockwise wild bootstrap under $H_0 : \rho(h) = 0$ for $h \geq 1$. The block size is $b_n = [c\sqrt{n}]$ with either $c = 1$ (i.e. fixed block size) or $c \sim U(0.5, 1.5)$ (i.e. randomized block size). Window size is $n \in \{240, 480, 720\}$. The number of bootstrap iterations is 5,000 for each window. Each point on the horizontal axis represents the window ID number. The shaded area represents nominal size $\alpha = 0.05$.

Figure 8: Sample Autocorrelations of Log Returns of Stock Price Indices



This figure plots sample autocorrelations at lags $1, \ldots, 25$ in full sample. The 95% confidence bands are constructed with Shao's (2011) blockwise wild bootstrap under the null hypothesis of white noise. The number of bootstrap samples is $M = 5000$. The black lines with "*" depict the sample autocorrelations, while the red lines with "x" depict the confidence bands.

References

Andrews, D. W. K., and W. Ploberger (1996): "Testing for Serial Correlation against an ARMA(1,1) Process," *Journal of the American Statistical Association*, 91, 1331–1342.

Chernozhukov, V., D. Chetverikov, and K. Kato (2013): "Gaussian Approximations and Multiplier Bootstrap for Maxima of Sums of High-Dimensional Random Vectors," *Annals of Statistics*, 41, 2786–2819.

Hansen, B. E. (1996): "Inference When a Nuisance Parameter Is Not Identified Under the Null Hypothesis," *Econometrica*, 64, 413–430.

Hill, J. B., and K. Motegi (2017): "A Max-Correlation White Noise Test for Weakly Dependent Time Series," Working Paper, Department of Economics at the University of North Carolina at Chapel Hill.

——— (2018): "Testing the White Noise Hypothesis of Stock Returns," Working Paper, Department of Economics at the University of North Carolina at Chapel Hill.

Hong, Y. (1996): "Consistent Testing for Serial Correlation of Unknown Form," *Econometrica*, 64, 837–864.

Lahiri, S. N. (1999): "Theoretical Comparisons of Block Bootstrap Methods," *Annals of Statistics*, 27, 386–404.

Liu, R. Y. (1988): "Bootstrap Procedures under some Non-I.I.D. Models," *Annals of Statistics*, 16, 1696–1708.

Nankervis, J. C., and N. E. Savin (2010): "Testing for Serial Correlation: Generalized Andrews-Ploberger Tests," *Journal of Business and Economic Statistics*, 28, 246–255.

——— (2012): "Testing for Uncorrelated Errors in ARMA Models: Non-Standard Andrews-Ploberger Tests," *Econometrics Journal*, 15, 516–534.

Politis, D. N., and J. P. Romano (1994): "The Stationary Bootstrap," *Journal of the American Statistical Association*, 89, 1303–1313.

Shao, X. (2011): "A Bootstrap-Assisted Spectral Test of White Noise under Unknown Dependence," *Journal of Econometrics*, 162, 213–224.

Wu, C. F. J. (1986): "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis," *Annals of Statistics*, 14, 1261–1295.

Xiao, H., and W. B. Wu (2014): "Portmanteau Test and Simultaneous Inference for Serial Covariances," *Statistica Sinica*, 24, 577–600.