Technometrics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/utch20

Scalable Methods for Multiple Time Series Comparison in Second Order Dynamics

Lei Jin & Bo Li

To cite this article: Lei Jin & Bo Li (20 Sep 2024): Scalable Methods for Multiple Time Series Comparison in Second Order Dynamics, Technometrics, DOI: 10.1080/00401706.2024.2388547

To link to this article: https://doi.org/10.1080/00401706.2024.2388547

TECHNOMETRICS
2024, VOL. 00, NO. 0, 1–15
https://doi.org/10.1080/00401706.2024.2388547

Taylor & Francis
Taylor & Francis Group

Scalable Methods for Multiple Time Series Comparison in Second Order Dynamics

Lei Jin[a] and Bo Li[b]

[a]Department of Mathematics and Statistics, Texas A&M University-Corpus Christi, Corpus Christi, TX; [b]Department of Statistics, University of Illinois Urbana-Champaign, Champaign, IL

ABSTRACT

Statistical comparison of multiple time series in their underlying frequency patterns has many real applications. However, existing methods are only applicable to a small number of mutually independent time series, and empirical results for dependent time series are only limited to comparing two time series. We propose scalable methods based on a new algorithm that enables us to compare the spectral density of a large number of time series. The new algorithm helps us efficiently obtain all pairwise feature differences in frequency patterns between M time series, which plays an essential role in our methods. When all M time series are independent of each other, we derive the joint asymptotic distribution of their pairwise feature differences. The asymptotic dependence structure between the feature differences motivates our proposed test for multiple mutually independent time series. We then adapt this test to the case of multiple dependent time series by partially accounting for the underlying dependence structure. Additionally, we introduce a global test to further enhance the approach. To examine the finite sample performance of our proposed methods, we conduct simulation studies. The new approaches demonstrate the ability to compare a large number of time series, whether independent or dependent, while exhibiting competitive power. Finally, we apply our methods to compare multiple mechanical vibrational time series.

ARTICLE HISTORY
Received May 2023
Accepted July 2024

KEYWORDS
Algorithm; Dynamics; Pairwise differences; Periodogram; Spectral method; Vibration data

1. Introduction

Mechanical systems, such as manufacturing machines, vehicles and their components, have significantly simplified and improved our lives. However, mechanical damage can occur from a variety of sources, which often causes unexpected mechanical breakdowns, financial losses, or even personnel casualties. Early mechanical damage detection is critical for preventing accidents and guaranteeing sufficient maintenance. It is shown that mechanical damage always results in changes in the frequency behaviors of vibration signals (Cempel and Tabaszewski 2007). As a result, vibrational data have been used as the basis for noninvasive damage detection. Mechanical damage can be detected by comparing the current signals to the reference vibrational signals obtained from a healthy system. To form "the reference database" for comparison, multiple vibrational time series from an undamaged system at various input force levels are often recorded (Sohn and Farrar 2001). A reliable reference database requires its participating vibrational time series to share common frequency behaviors, even at different input levels. However, this may not be true if the input levels are overly diversified, especially some of them being beyond the system's linear operating range (Karniel and Inbar 1999). Statistical methods are needed to evaluate if all members in a reference database exhibit the same frequency pattern.

The frequency pattern, or equivalently, the autocovariance structure, characterizes the second-order properties of a stationary time series. The comparison of time series in second-order dynamics has been widely studied under various contexts. Some of these methods were developed in the frequency domain by comparing frequency patterns, while others were developed in the time domain by comparing the autocovariance structure. Most existing methods in the literature, regardless of in frequency or time domain, were designed to compare only two time series. These methods include Coates and Diggle (1986), Diggle and Fisher (1991), Maharaj (2002), Alonso and Maharaj (2006), Lund, Bassily, and Vidakovic (2009), Dette and Paparoditis (2009), Decowski and Li (2015), Salcedo, Porto, and Morettin (2012), Jin and Wang (2016), Grant and Quinn (2017), Zhang and Tu (2018), Li and Lu (2018), Cirkovic and Fisher (2021), Jin (2021) and many others.

Methods for comparing multiple time series, however, are still scarce. Fokianos and Savvides (2008) proposed a likelihood ratio test on log-linear periodogram models to evaluate the equality of multiple spectral density functions. Their method requires choosing a time series as a baseline time series, so the results may rely on the choice of the baseline. Later, Jin (2015) developed testing procedures to compare correlation structures or the normalized spectral functions when these time series are independent of each other. However, their test seems to struggle with controlling the sizes when the number of time series, M, is not very small, for example, $M > 10$. Another approach to comparing the spectral densities of multiple time series can

CONTACT Lei Jin lei.jin@tamucc.edu Department of Mathematics and Statistics, Texas A&M University-Corpus Christi, Corpus Christi, TX 78412.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/TECH.

be seen in Jin (2018). All the above methods are developed for multiple time series that are independent of each other. Instead of comparing individual time series, Jin (2011) introduced a pre-planned contrast method to compare a group of time series to another group of time series. However, the method still compares two aggregated spectral densities, and the outcomes depend on the particular choice of these groups. Jentsch and Pauly (2015) and Zhang and Tu (2018) proposed methods for comparing multiple time series that may be dependent with each other, but both of them only showed simulation results of comparing two time series. Furthermore, results in Jentsch and Pauly (2015) appear to be sensitive to a bandwidth parameter involved in the method. Selecting multiple bandwidths to estimate a larger number of multivariate spectral and cross-spectral functions can be more challenging.

If the comparison procedure rejects the hypothesis that all time series have equal dynamics, a natural follow-up question to ask is which pairs of these time series are different. All the aforementioned methods themselves do not directly answer this question. Although Kalpakis, Gada, and Puttagunta (2001) proposed a distance measure of ARIMA time series based on the linear predictive coding cepstrum and Caiado, Crato, and Peña (2006) proposed an L_2 periodogram distance metric for time series classification, these distance metrics cannot tell if two time series are significantly different or not. In addition, it is unclear how these distance metrics work when these time series are not independent. For multiple independent time series, Fokianos and Savvides (2008) suggested performing pairwise comparisons among M time series with a Bonferroni correction. However, this would involve more than 1000 pairwise comparisons even when $M = 50$, and thus result in extensive computation. All these limitations call for new computationally efficient methods that can compare a large number of time series regardless of their inter-series dependency, and further provide details of the difference, if any.

We propose scalable methods to compare the frequency patterns of M time series for a large M such as $M = 100$. Our methods can evaluate if all M time series share the same frequency patterns, and meantime quantify all pairwise differences. We develop a new algorithm for our methods to obtain all pairwise feature differences between M time series that does not involve likelihood estimation, smoothing, and matrix inverse operations. The computation of the algorithm is only approximately equivalent to conducting M times matrix-vector multiplications. When all these time series are independent of each other, the joint asymptotic distribution for these pairwise feature differences is obtained. The asymptotic dependence structure between feature differences motivates our test for multiple independent time series and then we adapt the test to multiple dependent time series.

It is noteworthy that interests in multiple time series comparison are also prevalent in the data mining community beyond the statistical literature. Various data mining methods, such as matrix profile (Mercer and Keogh 2022; Der et al. 2022), and dynamic time warping (Chu et al. 2002; Alaee, Kamgar, and Keogh 2020) have been proposed. The similarity measured by those data mining methods is different from that of our proposed techniques. Both matrix profiles and dynamic time warping measure the distance between observations in the time domain to capture repeated patterns or functional structure such as motif, discord and joins in time series. In contrast, our methods evaluate differences in the second-order dynamics characterized by spectral density (or equivalently, autocovariance structures) between time series driven by random innovations. In addition, our methods are inference-based methods that can determine the significance of the differences in time series based on probability distributions.

The article is organized as follows. Section 2 introduces the proposed methods and develops the algorithm. A simulation study is presented in Section 3. In Section 4, we apply the proposed methods to compare multiple vibrational series of a mechanical system. Finally, a concluding remark is provided in Section 5.

2. Method

2.1. Hypotheses and Review of Periodogram

Consider multiple time series $\{X_{j,t}, j = 1, 2, \ldots, M; t = 1, 2, \ldots, T\}$, where j is the index of time series and t is the index of time. Each of these time series is assumed to be from a zero mean stationary random process. The autocovariance of the jth time series at lag h is defined as

$$\gamma_{j,h} = E(X_{j,t}X_{j,t+h}).$$

The autocovariance structure characterizes a stationary time series in the time domain. If the autocovariance $\gamma_{j,h}$, $h = 0, 1, 2, \ldots, T-1$ is absolutely summable, the spectral density of $X_{j,t}, t = 1, 2, \ldots, T$ is defined as

$$f_j(\omega) = \frac{1}{2\pi}\sum_{h\in\mathbb{Z}} \gamma_{j,h} e^{-\mathrm{i}h\omega} \tag{1}$$

for any real number ω, where i is the imaginary unit (see Brockwell and Davis 1991). The spectral density function describes the power distribution of a stationary time series in the frequency domain. The second-order dynamics of a stationary time series can be equivalently expressed either in its autocovariance structure or its spectral density.

Our main objective is to compare multiple time series in terms of their underlying models. One hypothesis for achieving this goal is to test whether the spectral densities of all M time series are the same. We then have the following null hypothesis:

$$H_0 : f_1(\omega) = f_2(\omega) = \cdots = f_M(\omega), \tag{2}$$

almost everywhere for $\omega \in (0, \pi)$. Sometimes, real-world problems such as damage detection of a mechanical system may desire to evaluate if the multiple time series have the same frequency pattern regardless of their magnitudes. In such cases, we can test whether their spectral densities are proportional to each other, or equivalently, whether their normalized spectral density functions are the same. The corresponding null hypothesis can be written as follows:

$$H_0^* : c_1 f_1(\omega) = c_2 f_2(\omega) = \cdots = c_M f_M(\omega), \tag{3}$$

where $c_1, c_2, \ldots, c_M$ are positive constants.

Many fundamental tools for spectral analysis are based on the periodogram, which allows us to identify the periodicity of

a time series and the relative strengths of these periodic components. The periodogram of the jth time series at the Fourier frequency $\omega_k = 2\pi k/T, k = 1, 2, \ldots, K = \lfloor T/2 \rfloor$ is defined as

$$I_j(\omega_k) = T^{-1} \left| \sum_{t=1}^{T} X_{j,t} \exp(-\mathrm{i}t\omega_k) \right|^2,$$

where $|x|$ denotes the modulus of x, and $\lfloor x \rfloor$ is the floor function of x. The lemma below describes some statistical properties of the periodogram.

Lemma 1. Suppose that

$$X_{j,t} = \sum_{l=-\infty}^{\infty} a_{j,l}\epsilon_{j,t-l}, \tag{4}$$

where $\{\epsilon_{j,t}\}$ is a sequence of identical and independent random variables with mean 0 and variance 1, $\sum_{l=-\infty}^{\infty} |a_{j,l}| < \infty$, and $E\epsilon_{j,1}^4 < \infty$ for all $j = 1, 2, \ldots, M$. For s distinct integers $1 \le k_1 < k_2 < \cdots < k_s \le K$, all of which may depend on T,

$$\left\{ \frac{I_j(\omega_{k_1})}{f_j(\omega_{k_1})}, \frac{I_j(\omega_{k_2})}{f_j(\omega_{k_2})}, \ldots, \frac{I_j(\omega_{k_s})}{f_j(\omega_{k_s})} \right\} \xrightarrow{d} \{E_{j,k}, 1 \le k \le s\},$$

as $T \to \infty$, where $E_{j,k}$, $k = 1, 2, \ldots s$, are independent and identically distributed standard exponential random variables, and "$\xrightarrow{d}$" stands for convergence in distribution.

This lemma is part of Theorem (10.3.2) in Brockwell and Davis (1991). The model in (4) is the general linear process of an infinite-order moving average process, forming a wide class of time series. Autoregressive moving average (ARMA) models, including the seasonal ARMA models in Kalpakis, Gada, and Puttagunta (2001), are all in this class. When data exhibit a nonperiodic mean trend, we can first use data preprocessing to remove the trend before applying methods like ours and Kalpakis, Gada, and Puttagunta (2001).

2.2. Statistical Tests

Similar to many previous approaches including Fokianos and Savvides (2008), our methods are developed based on the log ratio of periodograms. We define the log ratio between the periodograms of the ith and jth time series at the kth Fourier frequency as

$$\delta_{i,j,k} = \log(I_i(\omega_k)/I_j(\omega_k)),$$

where $k = 1, 2, \ldots, K$. To simplify the technical arguments, we assume that $|\delta_{i,j,k}|$ is bounded by a large number G_0, such as $G_0 = 10{,}000$, for all $1 \le i \ne j \le M$ and $k = 1, 2, \ldots, K$. After performing somewhat simple and not very strict calculations based on the limiting distributions of $\delta_{i,j,k}$, we find if both $f_i(\omega)$ and $f_j(\omega)$ are continuous and bounded away from zero, any finite sample with $|\delta_{i,j,k}| > G_0$ would provide extremely strong evidence to support $f_i(\omega) \ne f_j(\omega)$ for some interval of ω around ω_k. Therefore, there is no need for our methods to make decisions for a very large $|\delta_{i,j,k}|$.

The spectral function of a time series has infinite dimension, but its main feature can be quantified by a vector of dimension $R + 1$ for a small integer R via dimensional reduction. Consequently, the main feature of the difference between two spectral functions can also be summarized in an $R + 1$ dimensional vector. Shang (2014) discussed multiple approaches for functional dimensional reduction. One approach is to use basis function expansion involving expressing a function or a stochastic process as a linear combination of orthogonal basis functions. Suppose that $\mathbb{V} = \{v_0, v_1, v_2, \ldots\}$ is an orthonormal basis for the functional space of all continuous functions on interval $[0, 1]$, where $v_0 = 1$ and $\int_0^1 v_i^2(t)dt = 1$, for $i = 1, 2, \ldots$. Such basis functions can be obtained by appropriately centering and scaling a sequence of Legendre polynomials, a sequence of Fourier series, or other complete orthogonal systems.

For a positive integer r, let $\boldsymbol{\mathcal{V}}_r = \{\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_r\}$ where $\mathbf{v}_i = (v_i(\frac{\omega_1}{\pi}), v_i(\frac{\omega_2}{\pi}), \ldots, v_i(\frac{\omega_K}{\pi}))^{\mathrm{T}}$. For any positive integers i, j such that $1 \le i, j \le M$, we can model the first moment of the log periodogram ratio between the ith and jth time series as

$$E(\boldsymbol{\delta}_{i,j}) = \boldsymbol{\mathcal{V}}_R^T \boldsymbol{q}_{i,j}, \tag{5}$$

where $\boldsymbol{\delta}_{i,j} = (\delta_{i,j,1}, \delta_{i,j,2}, \ldots, \delta_{i,j,K})^{\mathrm{T}}$, and $\boldsymbol{q}_{i,j} = (q_{i,j,0}, q_{i,j,1}, \ldots, q_{i,j,R})^{\mathrm{T}}$. When $i = j$, which is the case of comparing the ith time series with itself, it is clear that $\boldsymbol{\delta}_{i,i} = \mathbf{0}_K$ and thus $\boldsymbol{q}_{i,j} = \mathbf{0}_{R+1}$, where $\mathbf{0}_K$ represents a zero vector of length K. Intuitively, if two time series $\{X_{i,t}\}$ and $\{X_{j,t}\}$ have the same underlying model, then asymptotically, $E(\boldsymbol{\delta}_{i,j})$ converges to $\mathbf{0}_K$, and thus $q_{i,j,k} = 0$ for $k = 0, 1, \ldots, R$. Hence, H_0 in (2) implies the simultaneous occurrence of $q_{i,j,l} = 0$ for all $1 \le i, j \le M; 0 \le l \le R$. When spectral densities $f_i(\omega)$ and $f_j(\omega_j)$ are proportional to each other, using Lemma 1 and the definition of $\delta_{i,j,k}$, it is straightforward to derive that $E(\boldsymbol{\delta}_{i,j})$ will asymptotically converge to a constant. Consequently, $q_{i,j,0}$ will no longer be asymptotically 0. However, when the normalized spectral densities of the two time series are the same, $q_{i,j,k}$ will still be asymptotically 0 for $k = 1, \ldots, R$. Therefore, H_0^* in (3) implies the simultaneous occurrence of $q_{i,j,l} = 0$ for all $1 \le i, j \le M; 1 \le l \le R$.

Let $\hat{\boldsymbol{q}}_{i,j} = (\hat{q}_{i,j,0}, \hat{q}_{i,j,1}, \ldots, \hat{q}_{i,j,R})^{\mathrm{T}} = \boldsymbol{\mathcal{V}}_R^{\mathrm{T}} \boldsymbol{\delta}_{i,j}$. This is shown in the Supplementary Materials to be asymptotically equivalent to the least squares estimate of $\boldsymbol{q}_{i,j}$. Compared to the latter, $\hat{\boldsymbol{q}}_{i,j}$ enjoys fast computation. We construct three tests based on $\hat{\boldsymbol{q}}_{i,j}$: one for multiple independent time series, one for multiple time series that may not be independent of each other, and a global test aimed at further improvement.

2.2.1. Test for Multiple Independent Time Series

To develop the test, we first study the asymptotic properties of $\hat{\boldsymbol{q}}_{i,j}$ for multiple independent time series. Let $\boldsymbol{I}_{k,k}$ be a k by k identity matrix for an integer k. The following theorem describes the asymptotic properties of $\hat{\boldsymbol{q}}_{i,j}$ under the H_0 defined in (2).

Theorem 2.1. Suppose we have M independent time series following (4). If their spectral densities are equal and bounded away from zero, then for any positive integers i and j, we have

$$\hat{\boldsymbol{q}}_{i,j} \xrightarrow{d} V_0(\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j),$$

as $T \to \infty$, where $\{\boldsymbol{\epsilon}_i, i = 1, 2, \ldots\}$ is a sequence of identical and independent multivariate Gaussian random vectors with

$E(\boldsymbol{\epsilon}_i) = \mathbf{0}_{R+1}$, $\text{cov}(\boldsymbol{\epsilon}_i) = \boldsymbol{I}_{R+1,R+1}$, and $V_0 = \sqrt{\text{var}(\log(\chi_2^2))}$, where χ_2^2 is a χ^2 random variable with 2 degrees of freedom.

The proof of Theorem 2.1 is deferred to the supplemental materials. Theorem 2.1 reveals the asymptotic normality and dependence structure of $\hat{\boldsymbol{q}}_{i,j}$ for different i and j, when all M time series are independent of each other. It can be seen that

$$\text{cov}(\hat{\boldsymbol{q}}_{i_1,j_1}, \hat{\boldsymbol{q}}_{i_2,j_2}) \to V_0^2 \text{cov}(\boldsymbol{\epsilon}_{i_1} - \boldsymbol{\epsilon}_{j_1}, \boldsymbol{\epsilon}_{i_2} - \boldsymbol{\epsilon}_{j_2}), \tag{6}$$

as $T \to \infty$, for any positive integers i_1, j_1, i_2 and j_2. If i_1, i_2, j_1 and j_2 are all different, the elements of covariance matrix $\text{cov}(\hat{\boldsymbol{q}}_{i_1,j_1}, \hat{\boldsymbol{q}}_{i_2,j_2})$ are all zero. Hence, most of the pairs between $\hat{\boldsymbol{q}}_{i_1,j_1}$ and $\hat{\boldsymbol{q}}_{i_2,j_2}$ such that $1 \le i_1 \ne j_1 \le M$ and $1 \le i_2 \ne j_2 \le M$ are independent. Under H_0, that is, the spectral density functions of M time series are equal, the asymptotic marginal distributions of $\hat{q}_{i,j,l}$ for all $1 \le i, j \le M$ and $0 \le l \le R$ are the same and all have mean 0. Any $\hat{q}_{i,j,l}$ greatly deviating from 0 may indicate rejection of the null hypothesis. We therefore propose the following test statistic for H_0 in (2):

$$\hat{S}_{Max,0} = \max_{1 \le i \le M; 1 \le j \le M; 0 \le l \le R} \hat{q}_{i,j,l}.$$

The null hypothesis is rejected if the test statistic is large.

We derive a slightly different test statistic for the H_0^* in (3) to determine whether the normalized spectral densities are equal or not. If $f_i(\omega)$ is proportional to $f_j(\omega)$, $q_{i,j,0}$ in (5) is determined solely by the ratio $f_i(\omega)/f_j(\omega)$, while $q_{i,j,k} = 0$ for $1 \le k \le R$. Intuitively, the value of the ratio $f_i(\omega)/f_j(\omega)$ is irrelevant to whether $f_i(\omega)$ is proportional to $f_j(\omega)$. Hence, $\hat{q}_{i,j,0}$ shall not be included in the statistic to test the equality of the normalized spectral densities. Let $\hat{\boldsymbol{q}}_{i,j}^*$ and $\boldsymbol{\epsilon}_i^*$ be the sub-vectors of $\hat{\boldsymbol{q}}_{i,j}$ and $\boldsymbol{\epsilon}$ without the corresponding first element, respectively. In fact, $\hat{\boldsymbol{q}}_{i,j}^* = (\hat{q}_{i,j,1}, \ldots, \hat{q}_{i,j,R})^{\mathrm{T}}$. Similar to Theorem 2.1, we have

$$\hat{\boldsymbol{q}}_{i,j}^* \xrightarrow{d} V_0(\boldsymbol{\epsilon}_i^* - \boldsymbol{\epsilon}_j^*),$$

when the normalized spectral densities of all time series are the same. We propose the following test statistic for H_0^* based on $\hat{\boldsymbol{q}}_{i,j}^*$:

$$\hat{S}_{Max,1} = \max_{1 \le i \le M; 1 \le j \le M; 1 \le l \le R} \hat{q}_{i,j,l}.$$

Both $\hat{S}_{Max,0}$ and $\hat{S}_{Max,1}$ can be expressed as $\hat{S}_{Max,k}$ with $k = 0$ and $k = 1$, respectively. By the continuous mapping theorem,

$$\hat{S}_{Max,k} \xrightarrow{d} \max_{1 \le i \le M; 1 \le j \le M; k \le l \le R} V_0(\epsilon_{i,l} - \epsilon_{j,l}),$$

as $T \to \infty$, where $\epsilon_{i,l}$ is the lth entry of $\boldsymbol{\epsilon}_i$ defined in Theorem 2.1. The analytic form of this limiting distribution may not be available. By the Borell-TIS inequality of Adler and Taylor (2007) and Lemma 2.1 of Chernozhukov, Chetverikov, and Kato (2013), an upper bound in the asymptotic probability of the test statistic above its expected value can be derived. But such results can only help obtain an upper limit of the p-value. Alternatively, the asymptotic critical values for $\hat{S}_{Max,k}$ may be obtained through simulating many Gaussian random variables as indicated by Theorem 2.1. However, for finite samples, the asymptotic critical values based on normal distributions may not work well.

We derive the critical value for $\hat{S}_{Max,k}$ from the perspective of multiple testing. Since each $\hat{q}_{i,j,l}$ of $\hat{S}_{Max,k}$ under H_0 asymptotically converges to a normal distribution with mean zero and a standard deviation of $\sqrt{2}V_0$, the test using $\hat{S}_{Max,k}$ is equivalent to testing if $q_{i,j,l} = 0$ via $|\hat{q}_{i,j,l}|$ simultaneously for all $1 \le i < j \le M$ and $k \le l \le R$ based on a common critical value. We further found a t distribution with degrees of freedom of $K - R + k$ works better than the asymptotic normal distribution for $\hat{q}_{i,j,l}/(\sqrt{2}V_0)$. Hence, we will use t distribution for each single test. Theorem 2.1 suggests that most of the individual tests are independent of each other, so a common critical value can be obtained via the Bonferroni adjustment. Because $\hat{q}_{ijl} = -\hat{q}_{jil}$, the total number of effective individual tests using $\hat{q}_{i,j,l}$ is $(R + 1 - k)M(M-1)/2$ instead of $M^2(R + 1 - k)$. Let $t_{\alpha,M,R,k}$ be the $1 - \alpha/((M-1)M(R+1-k))$ quantile of a t distribution with degrees of freedom $K - R + k$, where α is the significance level. Via the Bonferroni adjustment, the corresponding critical value for $\hat{S}_{Max,k}$ is $\sqrt{2}V_0 t_{\alpha,M,R,k}$.

2.2.2. Test for Multiple Dependent Time Series

When the M time series are dependent on each other and their underlying processes remain the same, we still seem to have $\hat{\boldsymbol{q}}_{i,j} \xrightarrow{d} V_0(\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_j)$ as $T \to \infty$. However, it is important to note that in this case, V_0 should be denoted as $V_{0,i,j}$, which may vary for different i and j. Additionally, $\{\boldsymbol{\epsilon}_i, i = 1, 2, \ldots\}$ may not be a sequence of identical and independent multivariate Gaussian random vectors. Both $V_{0,i,j}$ and the asymptotic dependence of $\hat{\boldsymbol{q}}_{i,j}$ depend on the unknown dependence among multiple time series. Effectively estimating and incorporating such unknown dependence into the asymptotic distribution of $\hat{S}_{Max,k}$ is challenging due to the significantly higher number of parameters involved, especially when M is large. Note that the dependence between the ith and jth time series has no effect on $E(\delta_{i,j,k})$, though it can have a significant effect on $\text{var}(\boldsymbol{\delta}_{i,j})$ and the asymptotic variance of $\hat{q}_{i,j,l}$. This is because $E(\delta_{i,j,k}) = E(\log(I_i(\omega_k)) - E(\log I_j(\omega_k)))$ does not involve the dependence between $I_i(\omega_k)$ and $I_j(\omega_k)$, while $\text{var}(\delta_{i,j,k})$ relies on their dependence. Fortunately, the sample standard deviation of $\boldsymbol{\delta}_{i,j}$ can be calculated easily and remains a consistent estimate for the standard error of $\boldsymbol{\delta}_{i,j}$, which automatically captures some information about the dependence between the ith and jth time series.

To partially compensate for the effects due to the unknown dependence, we adjust $\hat{S}_{Max,k}$ by using a standardized version of $\hat{q}_{i,j,l}$ in the test statistic. The new adjusted test statistic becomes

$$\hat{S}_{adj,k} = \max_{1 \le i \le M; 1 \le j \le M; k \le l \le R} \hat{t}_{i,j,l},$$

where $\hat{t}_{i,j,l} = \frac{\hat{q}_{i,j,l}}{s_{i,j}}$ and $s_{i,j}$ is the sample standard deviation of $\boldsymbol{\delta}_{i,j}$. When these time series are independent, we have

$$\hat{S}_{adj,k} \to \max_{1 \le i \le M; 1 \le j \le M; k \le l \le R} \frac{1}{\sqrt{2}}(\epsilon_{i,l} - \epsilon_{j,l}) \tag{7}$$

where $\epsilon_{i,l}$ is still the lth entry of $\boldsymbol{\epsilon}_i$ defined in Theorem 2.1 under the corresponding null. It is not difficult to see that $\frac{1}{\sqrt{2}}(\epsilon_{i,l} - \epsilon_{j,l})$ is a standard normal random variable for any $1 \le i \ne j \le M$ and $k \le l \le R$. However, when these

time series are not independent of each other, the result of (7) may not hold due to unknown correlation among $\hat{t}_{i,j,l}$, even if their asymptotic marginal distribution under the null is a standard normal. In this case, obtaining a critical value through an asymptotic distribution seems very challenging. Following the idea of obtaining the critical value for $\hat{S}_{Max,k}$, we can easily obtain the Bonferroni critical value for $\hat{S}_{adj,k}$ as $q_\alpha = t_{\alpha,M,R,k}$. As seen in the simulation study in Section 3.1, the test using $\hat{S}_{adj,k}$ and this critical value provides reasonable empirical Type I error rates when all M time series are moderately dependent with each other. Furthermore, we can determine if the ith and jth time series are significantly different at the significance level α by examining if

$$\hat{t}_{i,j,l} > q_\alpha,$$

for at least one l subject to $k \le l \le R$. Hence, our methods not only test if all M time series share frequency patterns, but also provide information about pairwise comparisons among these M time series.

2.2.3. A Global Test

The above tests using the maximum statistics can be very fast in computation and perform well for the time series with a large T. However, their finite sample performance may be less satisfactory and can be possibly improved. More importantly, these tests essentially employ a Bonferroni correction and may be conservative when the multiple tests are dependent. According to Sarkar and Chang (1997), the global test in Simes (1986) is more powerful than the classical Bonferroni adjustment while having the Type I error rate under control, when multiple test statistics are dependent but follow specific multivariate normal distributions. Since $\hat{q}_{i,j,l}$ for different integers $1 \le i \ne j \le M$ and $k \le l \le R$ approximately follows a multivariate normal distribution, we can adopt the global test approach in Simes (1986). The basic idea of the global test is still to evaluate

$$H_{0,ijl} : q_{i,j,l} = 0,$$

simultaneously for all i, j, l. For each $H_{0,ijl}$, we use the test statistic $\hat{t}_{i,j,l}$, because it partially accounts for the unknown dependence between the ith and jth time series. To obtain the p-value $p_{i,j,l}$ for $H_{0,ijl}$, we use the same t distribution with degrees of freedom $K-R+k$ as the reference distribution as in Section 2.2.1. Recall that $K = \lfloor T/2 \rfloor$. The p-value is twice of the probability that the reference distribution is above $|\hat{t}_{i,j,l}|$. We control the multiplicity of the p-values in the global test as follows. As mentioned earlier, due to the antisymmetric property $\hat{q}_{i,j,l} = -\hat{q}_{j,i,l}$, we only need to consider $H_{0,ijl}$ for $i < j$ and the total number of effective individual tests is $M(M-1)(R-k+1)/2$. We rank all p-values $p_{i,j,l}$, where $1 \le i < j \le M$ and $l = k, k+1, \ldots, R$, from smallest to largest. Let the corresponding order of $p_{i,j,l}$ be $L_{i,j,l}$. We reject H_0 if

$$p_{i,j,l} \le \frac{L_{i,j,l}\alpha}{(M-1)M(R+1-k)/2}$$

for at least one i, j and k subject to $1 \le i,j \le M$ and $k \le l \le R$.

2.3. A Scalable Algorithm to Calculate $\hat{q}_{i,j}$

All our tests depend on $\hat{\boldsymbol{q}}_{i,j} = (\hat{q}_{i,j,0}, \hat{q}_{i,j,1}, \ldots, \hat{q}_{i,j,R})^{\mathrm{T}}$, which is typically obtained through matrix-vector multiplication according to its definition and is our major computation demand. Though this computation is manageable, we develop a scalable algorithm to calculate $\hat{\boldsymbol{q}}_{i,j}$ even faster, to ensure our tests to be computationally efficient, especially when M is large. For M time series, there will be $M \times M$ pairwise comparisons on M log spectral functions, including M self-comparisons. We then introduce an array $\boldsymbol{Q}$ of dimension $M \times M \times (R+1)$ to store pairwise feature differences of M log spectral functions. We assign $\hat{q}_{i,j,k}$ to be the i, j, kth entry of the array $\boldsymbol{Q}$. The theorem below states that $\hat{\boldsymbol{q}}_{i,j}$ can be obtained via recursive vector additions instead of matrix-vector multiplication under certain conditions.

Theorem 2.2. For any positive integers $1 \le i < j \le M$, we have $\hat{\boldsymbol{q}}_{i,j} = -\hat{\boldsymbol{q}}_{j,i}$, and

$$\hat{\boldsymbol{q}}_{i,j} = \sum_{k=i+1}^{j} \hat{\boldsymbol{q}}_{k-1,k}.$$

The proof of Theorem 2.2 is given in the supplemental materials. According to Theorem 2.2, it is easy to see that $\hat{\boldsymbol{q}}_{i,j} = \sum_{k=i+1}^{j-1} \hat{\boldsymbol{q}}_{k-1,k} + \hat{\boldsymbol{q}}_{j-1,j} = \hat{\boldsymbol{q}}_{i,j-1} + \hat{\boldsymbol{q}}_{j-1,j}$, for any integers $1 \le i,j \le M$. If we have $\hat{\boldsymbol{q}}_{1,2}, \hat{\boldsymbol{q}}_{2,3}, \ldots, \hat{\boldsymbol{q}}_{M-1,M}$, we will be able to calculate any other $\hat{\boldsymbol{q}}_{i,j}$ recursively by vector additions instead of matrix-vector multiplications, for any integers $1 \le i,j \le M$. We develop Algorithm 1 to calculate $\boldsymbol{Q}$. Instead of calculating $\hat{\boldsymbol{q}}_{i,j}$ for all $1 \le i \le j \le M$ via $\boldsymbol{\mathcal{V}}_R^{\mathrm{T}}\boldsymbol{\delta}_{i,j}$, our algorithm only requires the calculation of M times matrix-vector multiplications plus $M(M-1)$ times vector additions. Thus, our method is much faster in computation, especially when M is large.

Input: Input M time series $\{X_{i,t}\}$, $i = 1, 2, \ldots, M$, and integer R
Output: A three dimensional array $\boldsymbol{Q}$
Set bound $G_0 = 10{,}000$, $R = 5$ and calculate matrix $\boldsymbol{\mathcal{V}}_R$
Define a three dimensional $M \times M \times (R+1)$ array $\boldsymbol{Q}$
Step 1: **foreach** $i = 1, 2, \ldots, M$, **do**
  Calculate the log periodogram $\log I_i(\cdot)$
Step 2: **foreach** $i = 1, 2, \ldots, M$, **do**
  Calculate the vector of log ratios $\boldsymbol{\delta}_{i,j}$, where $j = (i+1)\%M$
  Calculate $\hat{\boldsymbol{q}}_{i,j} = \boldsymbol{\mathcal{V}}_R^{\mathrm{T}}\boldsymbol{\delta}_{i,j}$
  Check if any elements of $|\boldsymbol{\delta}_{i,j,}|$ is above G_0 or not; if TRUE then set a flag.
Step 3: **foreach** $i = 1, 2, \ldots, M$, **do**
  $\boldsymbol{Q}[i, i,] = \boldsymbol{0}_{R+1}$
  $\boldsymbol{Q}[i, 1,] = \hat{\boldsymbol{q}}_{i,l}$, where $l = (i+1)\%M$
  foreach $j = i, \ldots, M$, **do**
    $\boldsymbol{Q}[i, j,] = \boldsymbol{Q}[i, j-1,] + \boldsymbol{Q}[j, 1,]$
    $\boldsymbol{Q}[j, i,] = -\boldsymbol{Q}[i, j,]$
return $\boldsymbol{Q}$.

Algorithm 1: Calculation for the three dimensional array $\boldsymbol{Q}$

Briefly, the computation complexity of our Algorithm is $O(MT\log T) + O(MT) + O(M^2)$. The term $O(MT\log T)$ arises

from the fast Fourier transformation of M time series of length T in Step 1, according to Winograd (1978). The term $O(MT)$ is primarily due to the M matrix-vector multiplications of Step 2, and the term $O(M^2)$ mainly accounts for the $\frac{M(M-1)}{2} - M$ addition operations of two vectors in Step 3.

Batista et al. (2014) discussed the invariances for time series distance measure. Since our methods are statistical inference-based, the distribution of each feature difference in $\boldsymbol{Q}$ is naturally invariant to different random errors in an asymptotic sense as long as the errors satisfy assumptions in Lemma 1. Additionally, due to the particular form of the statistics, our methods are also invariant to linear transformations including amplitude and offset. Specifically, our test statistics for evaluating the equality of multiple spectral densities remain the same when a common linear transformation is applied to all the time series in comparison. Furthermore, our test statistics for assessing the equality of normalized spectral densities remain the same even when each time series in comparison is transformed by a different linear transformation. Finally, our methods are based on the Fourier transformation which decomposes a time series into periodic components. Due to the orthogonality of Fourier basis, the proposed methods are also invariant to phase changes in any of these periodic components whose periodicity corresponds to a Fourier frequency.

3. Simulation Study

We conduct simulation studies to assess the finite sample performance of our proposed method under both null and alternative hypotheses. We use the Legendre polynomials with $R = 5$ as the orthogonal basis functions to construct (5). The choice of $R = 5$ is discussed and suggested by others such as Fokianos and Savvides (2008) and Jin (2021). The R function spec.pgram with the default cosine bell and tapering rate of $25/T$ was used to calculate the periodogram. We set the default level of significance to $\alpha = 0.05$ unless otherwise stated. For each scenario considered in the simulation, 1000 replications are performed.

To account for situations where these time series may be dependent with each other, we follow the below multivariate time series model to generate M individual time series:

$$\boldsymbol{X}_t = \boldsymbol{A}_1\boldsymbol{X}_{t-1} + \boldsymbol{B}_1\boldsymbol{z}_{t-1} + \boldsymbol{z}_t, \tag{8}$$

where

$$\boldsymbol{A}_1 = \begin{bmatrix} a_{11} & a_{12} & 0 & \dots & 0 & a_{1M} \\ a_{21} & a_{22} & a_{23} & \dots & 0 & 0 \\ \vdots & & & & & \vdots \\ a_{M1} & 0 & 0 & \dots & a_{M(M-1)} & a_{MM} \end{bmatrix},$$

$$\boldsymbol{B}_1 = \begin{bmatrix} b_{11} & b_{12} & 0 & \dots & 0 & b_{1M} \\ b_{21} & b_{22} & b_{23} & \dots & 0 & 0 \\ \vdots & & & & & \vdots \\ b_{M1} & 0 & 0 & \dots & b_{M(M-1)} & b_{MM} \end{bmatrix},$$

and $\boldsymbol{z}_t$ is a M-dimensional Gaussian innovation vector with $E(\boldsymbol{z}_t) = (0, 0, \dots, 0)^{\mathrm{T}}$ and $\mathrm{cov}(\boldsymbol{z}_t) = \boldsymbol{\Sigma}$. Let $\zeta = \min(|i - j|, (M + i - j))$ represent the index distance between the ith and the jth time series. This definition specifies that the index distance between the first and the Mth time series is 1 instead of $M - 1$, in a cyclic manner.

3.1. Size of the Tests

We examine the finite sample performance of the proposed tests when all M time series have equivalent underlying models. To simulate data under the null hypothesis, we impose constraints on the model coefficients and correlation matrix $\boldsymbol{\Sigma}$ in (8) to ensure that all individual models share the common model coefficients and the dependence between time series only depends on their index distance. Hence, we set that $a_{11} = a_{22} = \cdots = a_{MM}$, $a_{12} = a_{23} = \cdots = a_{(M-1)M} = a_{M1}$, $a_{21} = \cdots = a_{M(M-1)} = a_{1M}$, $b_{12} = b_{23} = \cdots = b_{(M-1)M} = b_{M1}$, $b_{21} = \cdots = b_{M(M-1)} = b_{1M}$, and $\Sigma_{ij} = \Sigma_{i'j'}$ if $j - i = j' - i'$ or $j - i = M - (j' - i')$ for $i < j, i' < j'$ and all $i, j, i', j' = 1, 2, \dots, M$, where Σ_{ij} is the (i, j)th entry of matrix $\boldsymbol{\Sigma}$. Basically, under these restrictions, each individual time series have the same form with a different time series index i.

We consider three groups of models (A, B, and C) for the null hypothesis, each with different inter-series dependence settings. Table 1 lists the specific model parameters for all models in each group. Group A includes models A1–A5, and each model generates M independent autoregressive–moving-average models with an autoregressive order of 1 and a moving-average order of 1, that is, ARMA(1,1) time series. By sharing common components for adjacent time series, the time series generated by each of the models B1–B5 in group B are dependent. The dependence between two time series decays as their index distance ζ increases. Group C introduces dependency through correlated innovations. Let $\boldsymbol{\Sigma}^*(\rho_1)$ be a correlation matrix such that $\Sigma^*_{ii}(\rho_1) = 1$ and $\Sigma^*_{ij}(\rho_1) = \rho_1$ if $i \neq j$. Each model in group C generates multiple dependent time series with pairwise correlation in their innovations as $\boldsymbol{\Sigma}^*(0.5)$, a moderate correlation strength. There are six models in group C. Models C1–C5 have the same model coefficients as Model A1–A5, but with different covariance matrices for innovations. We include Model C6 to represent a different scenario that combines a similar dependence structure employed in Group B with the innovation correlations used in the other models of Group C.

For each model in Table 1, we generate a multivariate time series with $M = 10, 25, 50, 100$ individual time series, respectively. Then we apply the proposed method to test if all M time series have the same underlying dynamics. The results are

Table 1. Model parameters of three groups of models for simulating multiple time series under the null hypothesis.

		a_{11}	a_{12}	a_{21}	b_{11}	b_{12}	b_{21}	$\boldsymbol{\Sigma}$
Group A	A1	0.0	0.0	0.0	0.0	0.0	0.0	$\boldsymbol{I}_M$
	A2	0.5	0.0	0.0	0.0	0.0	0.0	$\boldsymbol{I}_M$
	A3	0.0	0.0	0.0	0.5	0.0	0.0	$\boldsymbol{I}_M$
	A4	0.5	0.0	0.0	0.3	0.0	0.0	$\boldsymbol{I}_M$
	A5	0.7	0.0	0.0	2.0	0.0	0.0	$\boldsymbol{I}_M$
Group B	B1	0.3	0.15	0.15	0.0	0.0	0.0	$\boldsymbol{I}_M$
	B2	0.5	0.15	0.15	0.5	0.0	0.0	$\boldsymbol{I}_M$
	B3	0.0	0.25	0.25	0.5	0.0	0.0	$\boldsymbol{I}_M$
	B4	0.4	0.0	0.0	0.3	0.25	0.25	$\boldsymbol{I}_M$
	B5	0.6	0.1	0.1	2.0	0.25	0.25	$\boldsymbol{I}_M$
Group C	C1	0.0	0.0	0.0	0.0	0.0	0.0	$\boldsymbol{\Sigma}^*(0.5)$
	C2	0.5	0.0	0.0	0.0	0.0	0.0	$\boldsymbol{\Sigma}^*(0.5)$
	C3	0.0	0.0	0.0	0.5	0.0	0.0	$\boldsymbol{\Sigma}^*(0.5)$
	C4	0.5	0.0	0.0	0.3	0.0	0.0	$\boldsymbol{\Sigma}^*(0.5)$
	C5	0.7	0.0	0.0	2.0	0.0	0.0	$\boldsymbol{\Sigma}^*(0.5)$
	C6	0.1	0.25	0.25	0.0	0.0	0.0	$\boldsymbol{\Sigma}^*(0.5)$

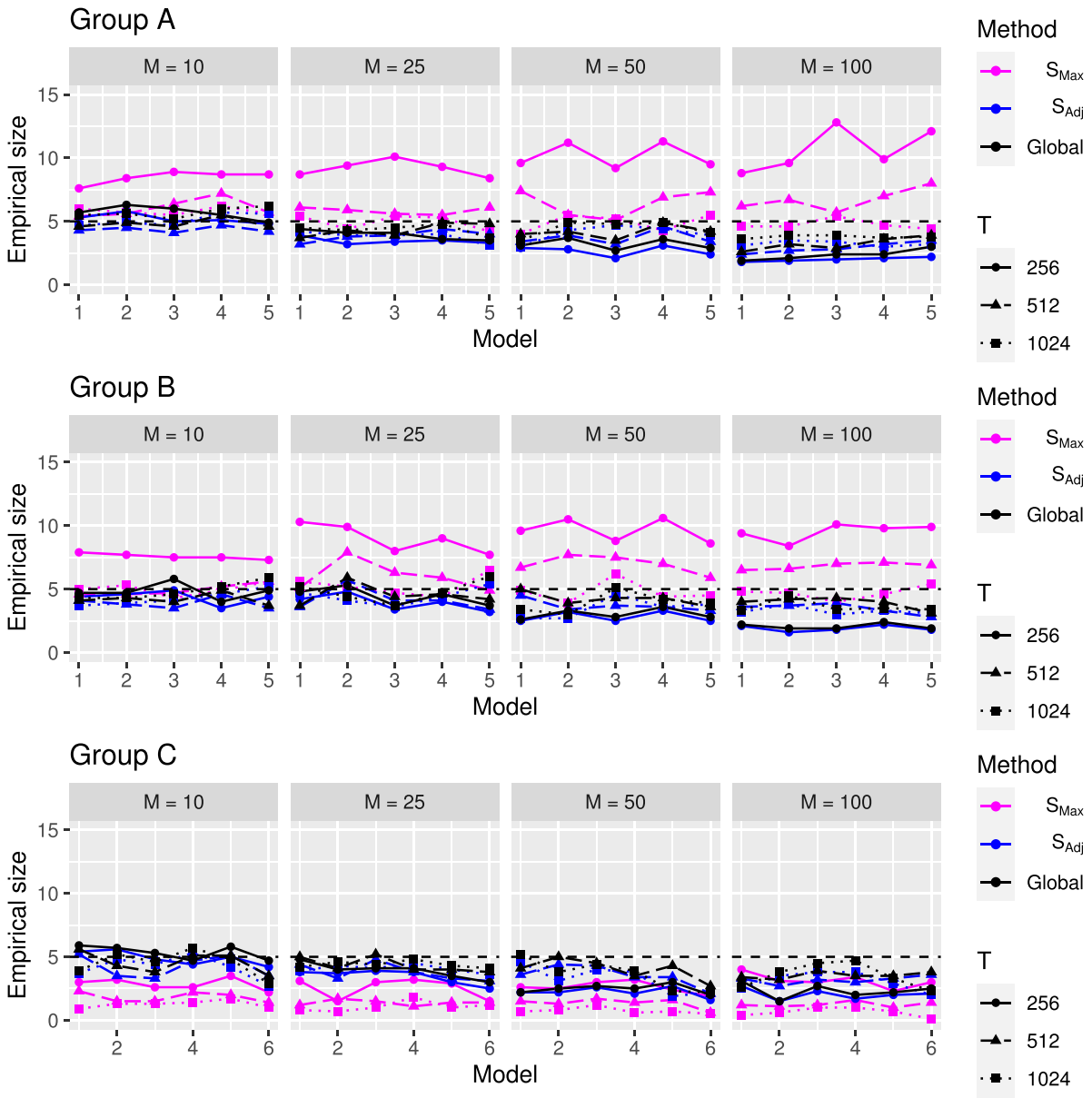

Figure 1. Empirical sizes for comparing M time series of length T. All time series are generated following the model specified in Table 1. In each plot, the horizontal dashed line indicates the nominal level of 0.05.

presented in Figure 1. When data are generated from groups A and B, it appears that the empirical Type I error rates of $\hat{S}_{\max,0}$ tend to be higher than the nominal level, especially at the combination of a small T and a large M, such as $T = 256$ and $M = 100$. The error rates for a small T deteriorate faster as M increases. However, when T is getting larger, all error rates become centered around or below the nominal level. For data generated by group C models, all the empirical rates are well controlled below the nominal level, though some cases seem to be slightly too conservative. These results indicate that $\hat{S}_{\max,0}$ may be sensitive to the dependency pattern among time series. The empirical Type I error rates of $\hat{S}_{adj,0}$ are well controlled across all scenarios. Their error rates when $M = 10$ are centered at the nominal level, but then tend to be below the nominal level, that is, the test becomes more conservative when M becomes larger. However, the error rates appear to approach the nominal level as T increases. Interestingly, the empirical Type I error rates are similar for all time series in different groups, regardless of whether the multiple time series are independent or not. This indicates that, as expected, $\hat{S}_{adj,0}$ is robust to the dependency structure among time series. The error rates of global test seem to be closer to the nominal level than those of $\hat{S}_{adj,0}$, though the global test still appears to be slightly conservative when T is small and M is large. Again, there is no clear distinction in the results when using the global test for the three groups of models.

3.2. Power of the Tests

This section investigates the finite sample performance of the proposed methods under the alternative hypothesis that the spectral densities of M time series are not all the same. Many previous studies, such as those by Lund, Bassily, and Vidakovic (2009) and Decowski and Li (2015), have all demonstrated the power of their methods by comparing a white-noise series to an

autoregressive (AR) time series with order 1, that is, AR(1). We modify their studies into two different settings for evaluating the power of our tests for multiple dependent time series. All settings described below are special cases of our general framework (8). Let $\Sigma^{\dagger}_{ij}(\rho_1) = \rho_1$ if $|i-j| < 2$, and $\Sigma^{\dagger}_{ij}(\rho_1) = 0$ otherwise. In setting I, for a given AR parameter ϕ, we generate the ith time series by following

$$X_{i,t} = (i-1)/(M-1)\phi X_{i,t-1} + z_{t,i}$$

where $z_{t,i}$ is the ith entry of innovation vector $\boldsymbol{z}_t$, and $\text{cov}(\boldsymbol{z}_t) = \boldsymbol{\Sigma}^{\dagger}(-0.25)$. This setting generates M time series from different AR models, with the innovations of two adjacent time series being negatively correlated. The dynamics of two time series apart from a smaller index distance ζ are more similar because their AR parameters are close. The spectral density of all these M time series is the same when $\phi = 0$. As ϕ increases, so does the difference in the spectral densities of these M time series. In setting II, we generate the first time series as $X_{1,t} = z_{t,i}$, and the rest $M-1$ time series as $X_{i,t} = \phi X_{i,t-1} + z_{t,i}$, for $1 < i < M$. The covariance between $z_{t,i}$ and $z_{t,j}$ follows $\boldsymbol{\Sigma}^{*}(0.5)$. This sets the model of the first time series different from the models of the rest $M-1$ time series. In this setting, the innovations of all M time series at the same time t have a correlation 0.5.

We additionally consider settings III and IV to compare multiple AR or MA time series, motivated by Lund, Bassily, and Vidakovic (2009) who compared an AR time series with a moving average (MA) time series. In these two settings, we generate the first $\lfloor M/2 \rfloor$ time series by an AR(1) process $X_{i,t} = \phi X_{i,t-1} + z_{i,t}$ when $i \leq \lfloor M/2 \rfloor$, and the rest time series by a MA(1) model $Y_{i,t} = \theta z_{i,t-1} + z_{i,t}$ when $i > \lfloor M/2 \rfloor$. The MA parameter $\theta = \text{sign}(\phi)\sqrt{\frac{\phi^2}{1-\phi^2}}$, where $\text{sign}(\phi) = 1$ if $\phi \geq 0$ and -1 otherwise. Setting III employs $\boldsymbol{\Sigma} = \boldsymbol{I}_M$ which means the M time series are mutually independent while setting IV employs $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{*}(0.5)$ which means the M time series are mutually dependent.

For each of the four settings described above, we generate $M = 10, 25, 50$, and 100 time series, respectively. We again apply the test using $\hat{S}_{\max,0}$, the test using $\hat{S}_{adj,0}$ and the global test using $\hat{t}_{i,j,l}$ all $1 \leq i \leq j \leq M$, $l = 0, 1, \ldots, R$ to all these time series. The results for settings I– IV are reported in Figures 2(a) and (b), 3(a) and (b), respectively. A common pattern for these plots is that the empirical power of all procedures increases either as ϕ increases or as T increases given ϕ is not tiny. Regarding the performance of different procedures, it appears that for setting I, the empirical power of $\hat{S}_{\max,0}$ is the highest for most scenarios. However, this is likely the spurious effect of its inflated size. Between the two other tests that can control their sizes below the nominal level, the global test seems to have a higher power in general. For settings II, the global test is still a winner. Settings III and IV reveals similar stories as setting I, and also conclude that the global test is generally more competitive. Note that the test using $\hat{S}_{adj,0}$ often performs very closely to the global test, especially when T is large.

The effect of M on powers appears to be heterogeneous across the four settings. When M increases, the power of all three tests seems to increase for settings I, III, and IV, while decreasing for setting II. Recall that setting II only sets the first time series different from the rest regardless of M. Intuitively, the power decrease for setting II makes sense because while many more comparisons will be involved when M increases there is only one different time series from the others. Still, the empirical power of all three tests reaches 100% when $\phi = 0.3$ and $T = 1024$ in setting II. This demonstrates that our proposed methods are capable of detecting a very small number of outlying time series hidden in a large number of background time series.

The methods proposed in Jin (2015) are able to test whether a small number of time series have the same normalized spectral densities, with the requirement that the multiple time series are mutually independent. We adopt a similar setup as in Jin (2015) to generate independent time series. More specifically, we generate time series $X_{i,t} = z_{t,i}$ if $i \leq \lfloor M/2 \rfloor$ and $X_{i,t} = \phi X_{i,t-1} + z_{t,i}$ if $i > \lfloor M/2 \rfloor$, and $\boldsymbol{\Sigma} = \boldsymbol{I}_M$. According to Jin (2015), different test statistics were studied, and it appears that the statistic $T_{M,\lfloor M/2 \rfloor}$ aligns with the current setting where $\lfloor M/2 \rfloor$ time series are distinct from the remaining ones. However, when T is not large and $M \geq 10$, $T_{M,\lfloor M/2 \rfloor}$ may suffer from inflated empirical Type I error rates. Therefore, we only compare our procedures to $T_{M,\lfloor M/2 \rfloor}$ when $M = 3, 5$, and 7. The results are presented in Figure 4(a). When $\phi = 0$, these M time series are from the same model, and thus its corresponding results represent the Type I error rates. It appears that the empirical Type I error rates for all tests across all scenarios are close to the nominal level. When $M = 3$, $T_{M,\lfloor M/2 \rfloor}$ seems to be more powerful than both $\hat{S}_{\max,1}$ and the global test. However, as M increases from 3 to 5 or 7, the performance of $T_{M,\lfloor M/2 \rfloor}$ deteriorates significantly. When $M = 5$ and $M = 7$, both $\hat{S}_{Max,1}$ and the global test for the equality of the normalized spectral densities, that is, H_0^*, are more powerful than $T_{M,\lfloor M/2 \rfloor}$, even under the settings restricted to meet the requirements of $T_{M,\lfloor M/2 \rfloor}$.

We additionally assess the performance of our proposed methods relative to two other approaches. Jin (2018) developed a test to assess if multiple independent time series have the same spectral densities. However, as noted by Jin (2018), the asymptotic null distribution of the test statistic is not sufficiently close to the empirical null distribution in practice, and hence Monte Carlo simulations were resorted to in order to obtain critical values varying by sample size and the number of time series. This leads to extensive computation even when M is small. On the other hand, Zhang and Tu (2018) developed a test to compare the spectra of two univariate time series that might be dependent. Though they mentioned a possible extension to compare multiple ($M > 2$) time series, both the procedure and the critical values for the extension are unclear. Hence, we perform the test by Zhang and Tu (2018) pairwise and then adjust the results by the Bonferroni correction. All the tuning parameters required to implement those two approaches follow the recommendations in their original papers. For these additional comparisons, we only consider the data generated by setting II with $M = 3, 5$, and 7. The simulation results are presented in Figure 4(b). The results corresponding to $\phi = 0$ represent the Type I error rates. It appears that the empirical Type I error rates for all tests are around the nominal level, but both proposed methods exhibit much higher power than the other two tests under this setting.

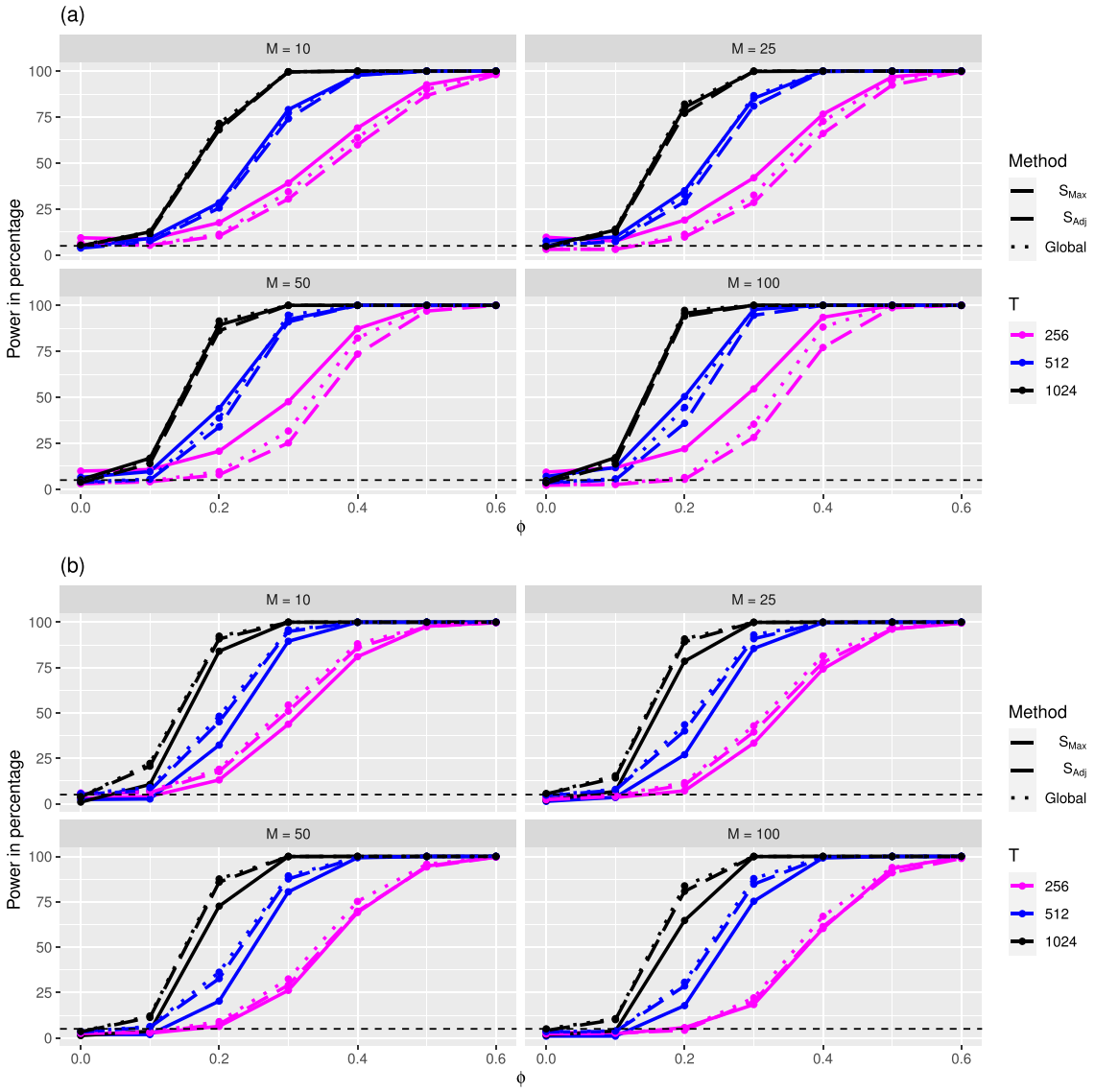

Figure 2. Empirical powers for comparing M time series of length T generated by the model specified in (a) Settings I and (b) Setting II, using different testing procedures. In each plot, $\phi = 0$ corresponds to the size and the horizontal dashed line indicates the nominal level of 0.05.

3.3. Robustness to Nonstationary Time Series

In reality, time series data often exhibit nonstationary behavior, so we now assess the effectiveness of our proposed methods when time series observations deviate from the stationary assumption. We introduce a time-varying function

$$s(t) = 2 + a \times \cos(t/512), \tag{9}$$

where a is a constant and $t = 1, 2, \ldots, T$. This function fluctuates periodically with a period of 512. Given a stationary time series X_t, a nonstationary time series Y_t can be generated by

$$Y_t = s(t)X_t. \tag{10}$$

The degree of non-stationarity increases as the constant a becomes larger. In our simulation, we set $a = 0.4$ which intuitively induces approximately $\pm 20\%$ fluctuation in the standard deviation of the time series $\{Y_t\}$ over time.

To study the empirical size of different tests for nonstationary time series, we generate Y_t using (10) with X_t following the models from Group C specified in Table 1. All three tests of $\hat{S}_{\max,0}$, $\hat{S}_{adj,0}$ and the global test are applied to the simulated nonstationary time series and the results are presented in Figure 5. It is seen that the empirical sizes of the proposed tests, especially the global test, align with the nominal level, suggesting robustness of our methods to nonstationary time series.

To study the empirical power of the proposed tests, we generate multivariate nonstationary time series with $M = 10$, 25, 50, and 100 using (10), where X_t follows setting I in Section 3.1 The results given in Figure 6 closely resemble those in Figure 2(a) for the stationary time series. This shows that our proposed tests are also robust to non-stationarity in terms of power.

3.4. Computing Time of the Proposed Methods

To illustrate the computational efficiency of our proposed methods and compare the computing time between the three tests for different M and T, we conduct a small experiment by generating M Gaussian white noises of length T and then applying each of the three tests to evaluate how M time series are different

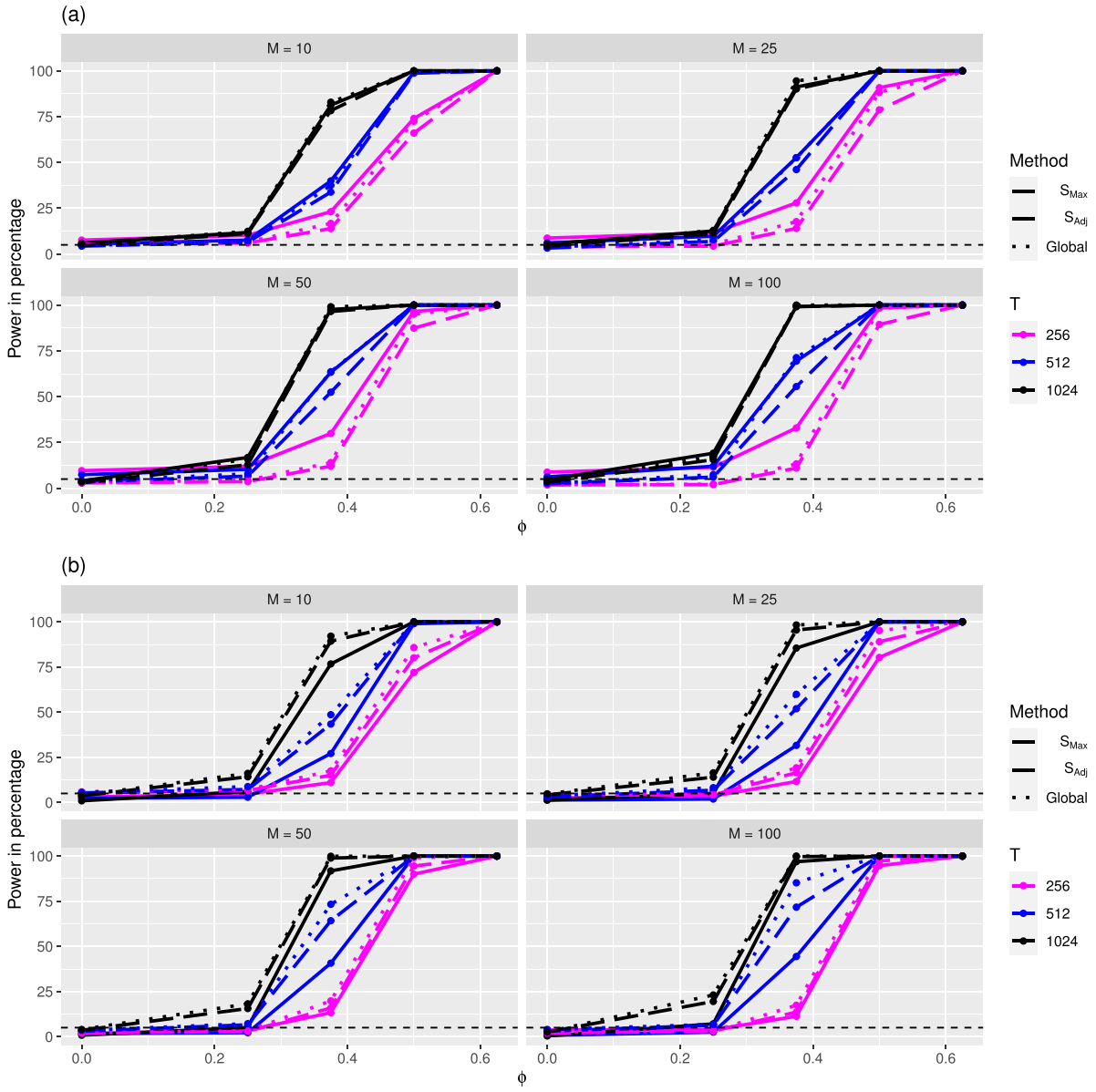

Figure 3. Empirical powers for comparing M time series of length T generated by the model specified in (a) Settings III and (b) Setting IV, using different testing procedures. In each plot, $\phi = 0$ corresponds to the size and the horizontal dashed line indicates the nominal level of 0.05.

from each other. Table 2 presents the average running times of different proposed procedures based on 100 simulation runs. The current R codes do not incorporate parallel computing. The computer for the simulation has a single I7-8565U CPU with a maximum speed of 4.60 GHz, and 16 GB of memory. The computation is very fast for all three procedures even when $M = 100$ and $T = 1024$. As T or M increases, the computational time also increases but at a different rate. It seems the computational cost is more sensitive to M. The running time of $S_{Max,0}$ is much lower than that of $S_{adj,0}$ and the global test, especially when M is large, such as $M = 100$. We also ran simulations to compare the computing time of our methods, Jin (2018) and Zhang and Tu (2018) in evaluating the difference between $M = 2$ time series. We found that the average running time of the proposed methods is at least 20 times faster than that of Jin (2018) and Zhang and Tu (2018). As M increases, the gap between the proposed methods and the existing methods using pairwise comparisons will become even wider.

4. Data Analysis

Damage in a mechanical system is generally detected by comparing the current vibrational signals collected from the system to the reference signals. We thus need to build a reference database that may consist of multiple vibrational signals recorded from the undamaged system at different input force levels/operation conditions (Sohn and Farrar 2001). To ensure the reliability of a reference database, it is necessary for the reference signals obtained from different input levels to have consistent frequency behaviors. We apply the proposed methods to evaluate if multiple time series in a reference database have the same normalized dynamics.

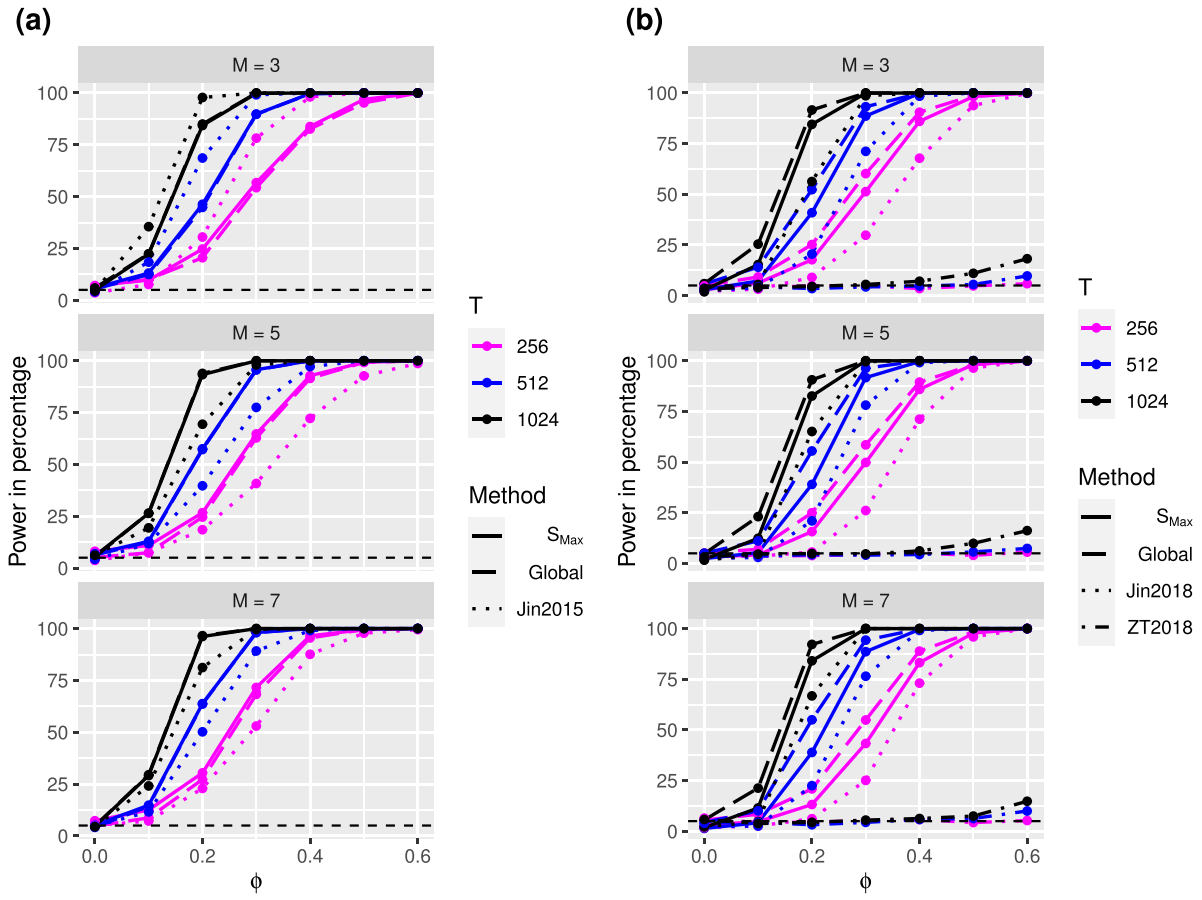

Figure 4. Comparing the empirical powers of the proposed methods and other tests at the significance level of 0.05. (a) $T_{M,\lfloor M/2 \rfloor}$ by Jin (2015) versus the proposed methods; (b) Tests by Jin (2018) and Zhang and Tu (2018) versus the proposed methods. "Jin2015", "Jin2018", and "ZT2018" represent methods of Jin (2015), Jin (2018), and Zhang and Tu (2018), respectively. In each plot, $\phi = 0$ corresponds to the size, and a horizontal dashed line represents the nominal level of 0.05.

Table 2. The average running time per run (in seconds) for different proposed procedures in 100 simulations when applied to compare M time series of length T.

	$\hat{S}_{Max,0}$			$\hat{S}_{adj,0}$			The global test		
M	T=256	T=512	T=1024	T=256	T=512	T=1024	T=256	T=512	T=1024
10	0.003	0.003	0.004	0.004	0.005	0.006	0.005	0.005	0.007
25	0.008	0.009	0.011	0.020	0.024	0.034	0.020	0.024	0.035
50	0.022	0.024	0.027	0.083	0.115	0.175	0.086	0.117	0.178
100	0.071	0.077	0.084	0.445	0.703	1.261	0.451	0.741	1.300

Twenty vibration signals were obtained from vibrational data sets collected by attaching an electro-dynamic shaker to a three-story frame structure in a laboratory experiment. The shaker was propelled by a random waveform with a uniform energy distribution within the frequency range of 0 to 200 Hz. See Fasel et al. (2003) for more details. The vibration data, driven by exogenous input, are often modeled via autoregressive-exogenous (ARX) (p, q) models, assuming that the current system output is a linear combination of the preceding p system outputs and the preceding q system inputs. As shown in Roy, Bhattacharya, and Ray-Chaudhuri (2015), the coefficients of ARX models are determined by system physical characteristics, such as structural stiffness and mass. Regarding our data, the random waveform input plays the role of the random innovations and the physical characteristics of the structure determine the ARX coefficients. The ARX coefficients, in turn, determine the temporal dependency structure (auto-covariance structure) of the ARX time series. Any change, such as damage in the physical characteristics of the frame structure, will alter the ARX coefficients and thus the temporal dependence structure. Various input voltage values were applied to power the shaker. The specific input voltage was 0.075v for the first five time series (group 1), 0.128v for the next five time series (group 2), then 0.25v for the following five time series (group 3), and finally 1.0v for the last five time series (group 4). Each signal has a length of 512, over 1 second at a rate of 512 Hz. Figure 7 displays these time series in a row order, that is, each row represents a different group. All time series were obtained when the system was healthy.

To compare these 20 time series, we calculate the proposed test statistics $\hat{S}_{\max,0}$, $\hat{S}_{adj,0}$, $\hat{S}_{\max,1}$ and $\hat{S}_{adj,1}$, with $R = 5$. The results are $\hat{S}_{\max,0} = 81.10$, $\hat{S}_{adj,0} = 44.92$, $\hat{S}_{\max,1} = 8.43$ and $\hat{S}_{adj,1} = 4.43$. The critical values for these statistics at the 0.05 significance level are 7.54, 4.16, 7.46, and 4.11, respectively. Recall that the statistics $\hat{S}_{\max,0}$ and $\hat{S}_{adj,0}$ test whether the spectral

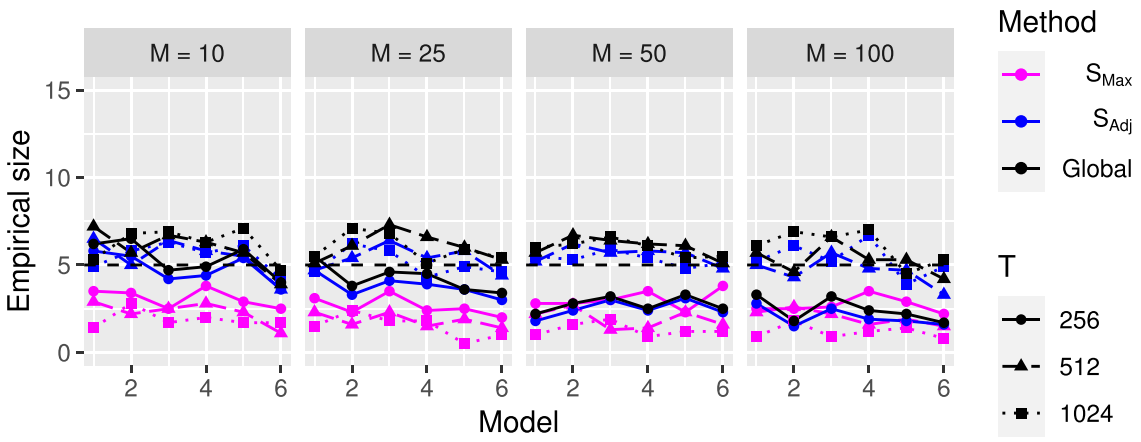

Figure 5. Empirical sizes for comparing M nonstationary time series of length T. The horizontal dashed line indicates the nominal level of 0.05.

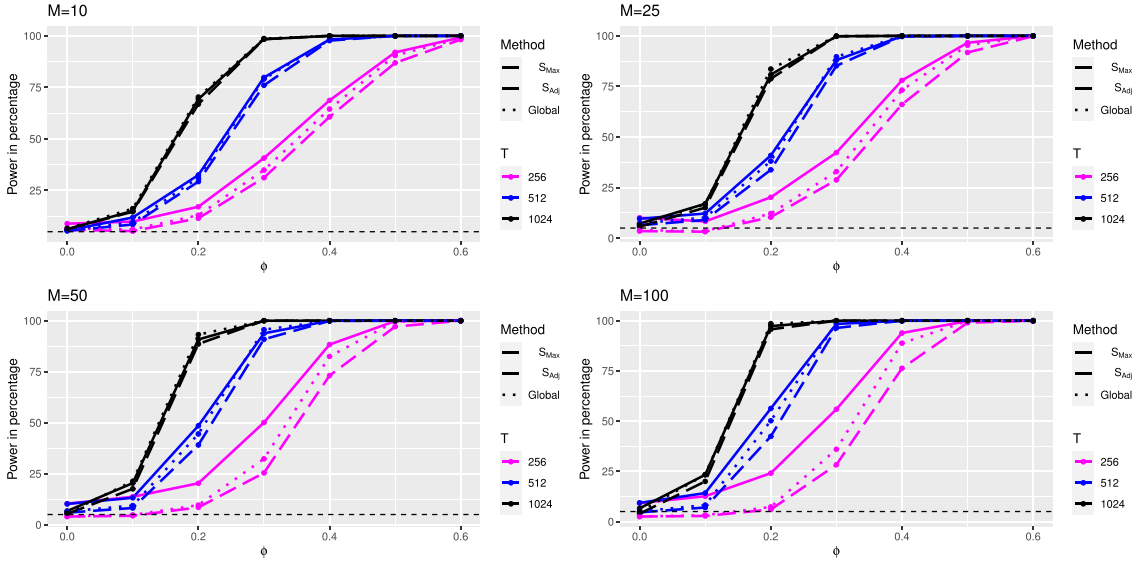

Figure 6. Empirical powers for comparing M nonstationary time series of length T using different testing procedures. In each plot, $\phi = 0$ corresponds to the size and the horizontal dashed line indicates the nominal level of 0.05.

densities of multiple time series are equal, while $\hat{S}_{\max,1}$ and $\hat{S}_{adj,1}$ test if the normalized spectral densities are equal. We conclude that the spectral densities of these signals are not all the same because both $\hat{S}_{\max,0}$ and $\hat{S}_{adj,0}$ are much higher than their corresponding critical values. For the normalized spectral densities, both $\hat{S}_{adj,1}$ and $\hat{S}_{\max,1}$ are slightly above their corresponding critical values. These results indicate that the signals collected at different inputs have different magnitudes. Their normalized frequency patterns are similar but may not be identical, suggesting that the dynamics of the system may vary slightly with different input levels. Via the feature differences in the array $\boldsymbol{Q}$, it is easy to see how the signals differ from each other. Figure 8 presents the result of the pairwise comparison. There is a clear pattern in Figure 8(a). It shows that the time series with the same input level have identical frequency densities, while the time series with different input levels have different spectral densities. According to Figure 8(b), the normalized spectral densities of all time series except the 14th are the same. By checking the values of $\boldsymbol{Q}$ in detail, we found only three quantities in $\boldsymbol{Q}$ related to the 14th time series slightly above the critical values, indicating that the differences are not very significant. Still, we should be cautious to use these reference signals collected at very different voltage inputs for damage detection.

We applied the tests developed in Jin (2018) and Zhang and Tu (2018) to assess whether the spectral densities of these signals are identical. At a significance level of 0.05, both tests reached the same conclusion that the spectral densities of all these 20 time series are not identical, consistent with the findings of the proposed methods. However, the test of Jin (2018) is unable to further identify which specific time series differ from one another. The pairwise comparison based on Zhang and Tu (2018) reveals a pattern (omitted here) very different from the expectation that the spectral densities of time series within the same voltage input group are likely identical. Our results in Figure 8(a), however, align well with this expectation by showing a block pattern.

Some mechanical systems have inherent repeated pattern/shape in vibrational signals. Those patterns can be altered by damage. For such case, Mercer and Keogh (2022) introduced the novelets method to detect possible damages by identifying emerging patterns, which turns to be very useful for industrial process monitoring. The dynamic wrapping

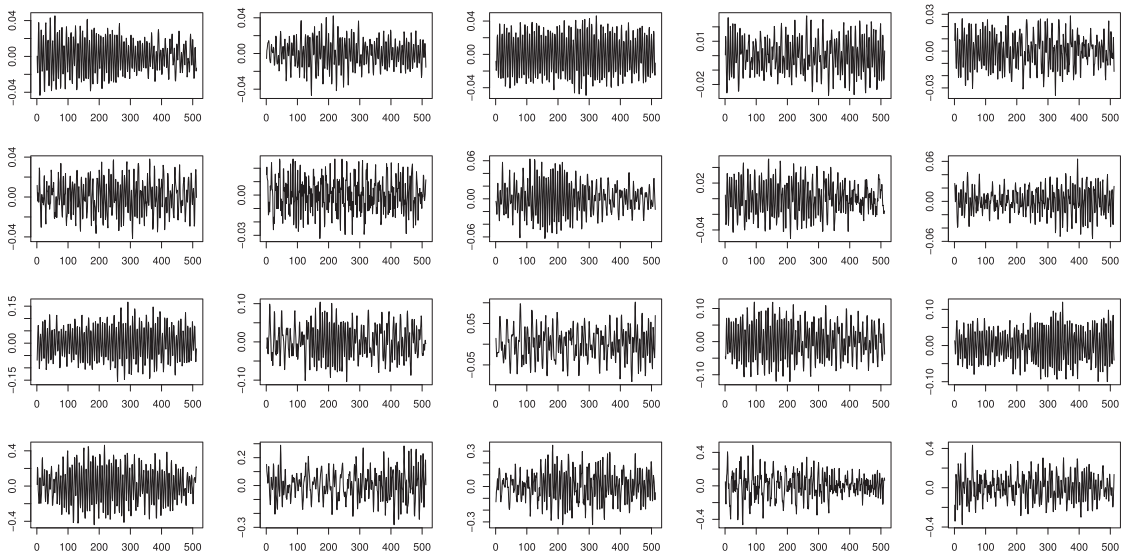

Figure 7. Twenty vibration signals from a laboratory experiment under four different input levels.

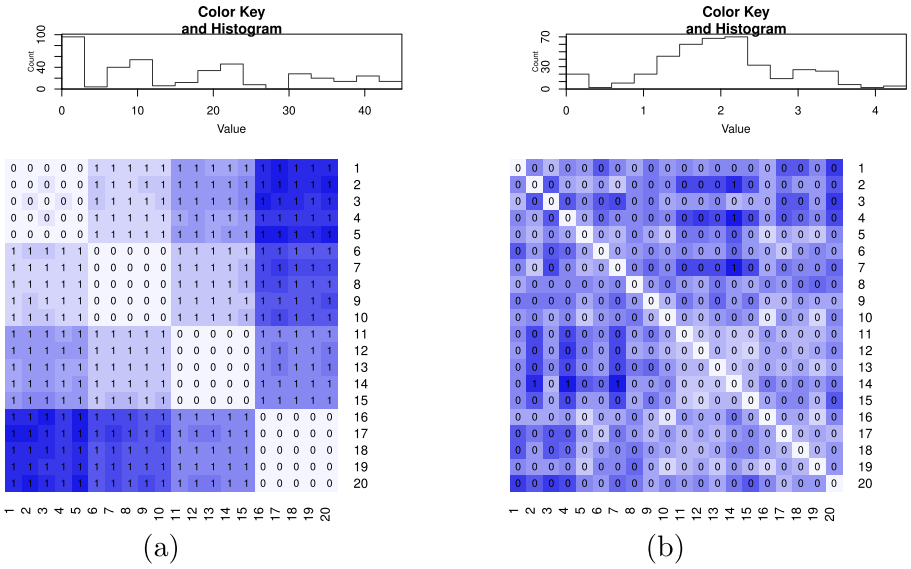

Figure 8. Results of pairwise comparison of 20 time series in terms of (a) spectral densities and (b) normalized spectral densities. Color indicates the value of $\max_l \hat{q}_{ijl}$, the maximum feature difference between the ith and jth time series. Cells with "1" indicate that differences between the corresponding time series are significant. The indices of the 20 time series are provided on both the right-hand and bottom sides of the plot. A histogram of the maximum feature difference is overlaid onto the legend on top of each plot.

method is another effective approach for detecting differences in repeated patterns between two signals when there is little to no noise in the observations. However, in the presence of non-negligible noise, it is challenging to use the dynamic wrapping distance to decide whether two processes have the same repeated pattern. In contrast to their applications, our vibrational data were generated with a random waveform input, which will not lead to repeated patterns in the vibrational signals as those in Mercer and Keogh (2022). Hence, under the current or similar settings, our proposed methods are more suitable.

5. Conclusion

We proposed computationally efficient methods to formally evaluate the significance of the differences between a large number of time series in terms of frequency patterns. According to Cai and Sun (2011), the emerging large-scale hypothesis testing that may consist of thousands or more simultaneous tests poses many challenges not present in smaller-scale studies. Our methods are constructed based on pairwise feature differences between M time series, for which we developed a computationally efficient algorithm to ensure its scalability. Previous literature only showed results comparing two dependent time series or comparing a few independent time series. We have demonstrated that our methods can be applied to a much larger number of time series, such as $M = 100$. Our algorithm has reduced the main computational complexity of obtaining feature differences for pairwise comparisons from $O(M^2)$ to $O(M)$. As shown in Table 2, it takes only around 0.084 sec to complete about 5000 times pairwise comparisons for $M = 100$ time series of length $T = 1024$ using $\hat{S}_{Max,0}$. This shows that our methods,

especially the one with $\hat{S}_{Max,0}$, can process well beyond 100 time series within an acceptable amount of time. Without modeling the unknown dependence structure among multiple time series, our proposed approaches work well in comparing multiple time series that may or may not be independent of each other. In addition, our proposed methods allow us to identify which pairs of time series are different without much additional effort. They appear to be useful in addressing real-world problems.

The asymptotic distribution of the pairwise feature differences exhibits an interesting dependence structure. When the time series are independent of each other, most $\hat{q}_{ijk}$, $1 \le i \ne j \le M$, $k = 0, 1, \ldots, R$ are independent which allows us to easily obtain a working critical value for our test statistic $\hat{S}_{Max,k}$ via a t distribution with the Bonferroni correction. Since the test statistic $\hat{S}_{Max,k}$ can be computed very efficiently even for a large T, we recommend this test when the multiple time series are independent of each other and T is large. In contrast, both $\hat{S}_{adj,k}$ and the global test are more suitable for dependent scenarios and tend to yield better results. Among the three methods, the global test often exhibits the highest power. We observe that some of our tests may be slightly conservative, especially when M is large and T is small. Possible solutions to address this issue are to generate critical values through simulations of the procedures or to apply finite sample adjustments.

Supplementary Materials

The supplementary materials contain (A) proof of theorems, (B) empirical studies to compare the proposed method with dynamic time warping, and (C) the R code for replicating Figures 1–8 in this article and Figures 1, 3–5 of the supplement.

Acknowledgments

The authors wish to thank the editor, the associate editor and two anonymous referees for their valuable comments and suggestions which substantially improved an earlier version of the article.

Disclosure Statement

No potential conflict of interest was reported by the authors.

Funding

Li's research is partially supported by NSF-DMS-2124576.

References

Adler, R. J., and Taylor, J. E. (2007), "Gaussian Inequalities," *Random Fields and Geometry*, 49–64. [4]

Alaee, S., Kamgar, K., and Keogh, E. (2020), "Matrix Profile XXII: Exact Discovery of Time Series Motifs under DTW," in *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 900–905. IEEE. [2]

Alonso, A., and Maharaj, E. A. (2006), "Comparison of Time Series Using Subsampling," *Computational Statistics & Data Analysis*, 50, 2589–2599. [1]

Batista, G. E., Keogh, E. J., Tataw, O. M., and De Souza, V. M. (2014), "CID: An Efficient Complexity-Invariant Distance for Time Series," *Data Mining and Knowledge Discovery*, 28, 634–669. [6]

Brockwell, J. P., and Davis, A. R. (1991), *Time Series: Theory and Methods* (2nd ed.), New York: Springer. [2,3]

Cai, T. T., and Sun, W. (2011), "A Compound Decision-Theoretic Approach to Large-Scale Multiple Testing," in *High-dimensional Data Analysis*, pp. 75–116, World Scientific. [13]

Caiado, J., Crato, N., and Peña, D. (2006), "A Periodogram-based Metric for Time Series Classification," *Computational Statistics & Data Analysis*, 50, 2668–2684. [2]

Cempel, C., and Tabaszewski, M. (2007), "Multidimensional Condition Monitoring of Machines in Non-Stationary Operation," *Mechanical Systems and Signal Processing*, 21, 1233–1241. [1]

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013), "Gaussian Approximations and Multiplier Bootstrap for Maxima of Sums of High-Dimensional Random Vectors," *The Annals of Statistics*, 41, 2786–2819. [4]

Chu, S., Keogh, E., Hart, D., and Pazzani, M. (2002), "Iterative Deepening Dynamic Time Warping for Time Series," in *Proceedings of the 2002 SIAM International Conference on Data Mining*, pp. 195–212, SIAM. [2]

Cirkovic, D., and Fisher, T. J. (2021), "On Testing for the Equality of Autocovariance in Time Series," *Environmetrics*, 32, e2680. [1]

Coates, D. S., and Diggle, P. (1986), "Tests for Comparing Two Estimated Spectral Densities," *Journal of Time Series Analysis*, 7, 7–20. [1]

Decowski, J., and Li, L. (2015), "Wavelet-based Tests for Comparing Two Time Series with Unequal Lengths," *Journal of Time Series Analysis*, 36, 189–208. [1,7]

Der, A., Yeh, C. M., Wu, R., Wang, J., Zheng, Y., Zhuang, Z., Wang, L., Zhang, W., and Keogh, E. (2022), "Matrix Profile XXVII: A Novel Distance Measure for Comparing Long Time Series," in *2022 IEEE International Conference on Knowledge Graph (ICKG)*, pp. 40–47, IEEE. [2]

Dette, H., and Paparoditis, E. (2009), "Bootstrapping Frequency Domain Tests in Multivariate Time Series with An Application to Comparing Spectral Densities," *Journal of the Royal Statistical Society, Series B*, 71, 831–857. [1]

Diggle, P. J., and Fisher, N. I. (1991), "Nonparametric Comparison of Cumulative Periodograms," *Applied Statistics*, 40, 423–434. [1]

Fasel, T. R., Sohn, H., Park, G., and Farrar, C. R. (2003), "Application of Frequency Domain ARX Models and Extreme Value Statistics to Impedance-based Damage Detection," in *ASME International Mechanical Engineering Congress and Exposition* (Vol. 37076), pp. 289–297. [11]

Fokianos, K., and Savvides, A. (2008), "On Comparing Several Spectral Densities," *Technometrics*, 50, 317–331. [1,2,3,6]

Grant, A. J., and Quinn, B. G. (2017), "Parametric Spectral Discrimination," *Journal of Time Series Analysis*, 38, 838–864. [1]

Jentsch, C., and Pauly, M. (2015), "Testing Equality of Spectral Densities Using Randomization Techniques," *Bernoulli*, 21, 697–739. [2]

Jin, L. (2011), "A Data-Driven Test to Compare Two or Multiple Time Series," *Computational Statistics & Data Analysis*, 55, 2183–2196. [2]

——— (2015), "Comparing Autocorrelation Structures of Multiple Time Series via the Maximum Distance between Two Groups of Time Series," *Journal of Statistical Computation and Simulation*, 85, 3535–3548. [1,8,11]

——— (2018), "A Frequency-Domain Test to Check Equality in Spectral Densities of Multiple Time Series with Unequal Lengths," *Journal of Time Series Analysis*, 39, 618–633. [2,8,10,11,12]

——— (2021), "Robust Tests for Time Series Comparison based on Laplace Periodograms," *Computational Statistics & Data Analysis*, 160, 107223. [1,6]

Jin, L., and Wang, S. (2016), "A New Test for Checking the Equality of the Correlation Structures of Two Time Series," *Journal of Time Series Analysis*, 37, 335–368. [1]

Kalpakis, K., Gada, D., and Puttagunta, V. (2001), "Distance Measures for Effective Clustering of Arima Time-Series," in *Proceedings 2001 IEEE International Conference on Data Mining*, pp. 273–280, IEEE. [2,3]

Karniel, A., and Inbar, G. F. (1999), "Linear Systems description," in *Modern Techniques in Neuroscience Research*, pp. 589–625, Berlin: Springer. [1]

Li, L., and Lu, K. (2018), "Tests for the Equality of Two Processes' Spectral Densities with Unequal Lengths Using Wavelet Methods," *Journal of Time Series Analysis*, 39, 4–27. [1]

Lund, R., Bassily, H., and Vidakovic, B. (2009), "Testing Equality of Stationary Autocovariances," *Journal of Time Series Analysis*, 30, 332–348. [1,7,8]

Maharaj, E. A. (2002), "Comparison of Non-stationary Time Series in the Frequency Domain," *Computational Statistics & Data Analysis*, 40, 131–141. [1]

Mercer, R., and Keogh, E. (2022), "Matrix Profile xxv: Introducing Novelets: A Primitive that Allows Online Detection of Emerging Behaviors in Time Series," in *2022 IEEE International Conference on Data Mining (ICDM)*, pp. 338–347. IEEE. [2,12,13]

Roy, K., Bhattacharya, B., and Ray-Chaudhuri, S. (2015), "ARX Model-based Damage Sensitive Features for Structural Damage Localization Using Output-Only Measurements," *Journal of Sound and Vibration*, 349, 99–122. [11]

Salcedo, G. E., Porto, R. F., and Morettin, P. A. (2012), "Comparing Non-stationary and Irregularly Spaced Time Series," *Computational Statistics & Data Analysis*, 56, 3921–3934. [1]

Sarkar, S. K., and Chang, C. (1997), "The Simes Method for Multiple Hypothesis Testing with Positively Dependent Test Statistics," *Journal of the American Statistical Association*, 92, 1601–1608. [5]

Shang, H. L. (2014), "A Survey of Functional Principal Component Analysis," *AStA Advances in Statistical Analysis*, 98, 121–142. [3]

Simes, R. J. (1986), "An Improved Bonferroni Procedure for Multiple Tests of Significance," *Biometrika*, 73, 751–754. [5]

Sohn, H., and Farrar, C. R. (2001), "Damage Diagnosis Using Time Series Analysis of Vibration Signals," *Smart Materials and Structures*, 10, 446. [1,10]

Winograd, S. (1978), "On Computing the Discrete Fourier Transform," *Mathematics of Computation*, 32, 175–199. [6]

Zhang, S., and Tu, X. (2018), "Tests for Comparing Time-Invariant and Time-Varying Spectra based on the Pearson Statistic," *Journal of Time Series Analysis*, 39, 709–730. [1,2,8,10,11,12]