Journal of Statistical Planning and Inference 211 (2021) 241–255

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Testing for adequacy of seasonal adjustment in the frequency domain

Tucker McElroy [a,*], Anindya Roy [b]

[a] Research and Methodology Directorate, U.S. Census Bureau, 4600 Silver Hill Road, Washington, D.C. 20233-9100, United States of America
[b] University of Maryland Baltimore County, United States of America

ARTICLE INFO

Article history:
Received 26 November 2019
Received in revised form 23 June 2020
Accepted 27 June 2020
Available online 3 July 2020

Keywords:
Fixed-b asymptotics
Spectral peaks
Visual significance

ABSTRACT

Peaks in the spectral density estimates of seasonally adjusted data are indicative of an inadequate adjustment. Spectral peaks are currently assessed in the X-13ARIMA-SEATS program via the visual significance (VS) approach; this paper provides a rigorous statistical foundation for VS by defining measures of uncertainty for spectral peak measures, allowing for formal hypothesis testing, using the framework of fixed-bandwidth fraction asymptotics for taper-based spectral density estimates. The simulation results show that the test has good size and power properties for a variety of peak features.

Published by Elsevier B.V.

1. Introduction

Quarterly or monthly economic time series typically exhibit seasonality, most often described via a nonstationary stochastic process with unit root corresponding to the known seasonal frequencies (Bell and Hillmer, 1983); also see the discussion in Chapter 3 of Hylleberg (1986). Adequate estimation and removal of seasonality should correspond to the absence of spectral peaks at these same seasonal frequencies in the adjusted series. If there is residual seasonality present, the adjustment is inadequate, and hence testing for residual seasonality is a problem of widespread importance; millions of time series are seasonally adjusted each month at statistical agencies around the world, many of whom utilize the software program X-13ARIMA-SEATS (U.S Census Bureau, 2015) of the U.S. Census Bureau. Testing for residual seasonality involves data arising from time series that typically have trend nonstationarity, but no longer have seasonal nonstationarity, and therefore applications of the tests can be formulated for processes that are stationary after differencing. This is the case considered in this paper.

One method of testing for the presence of residual seasonality is based on detection of spectral peaks (Findley, 2005). The detection of seasonality in raw, or unadjusted data, is a different problem that historically has used different tools — we provide a brief discussion in order to frame the problem of testing for residual seasonality.

One approach to the detection of seasonality (in raw data) is to postulate as a null hypothesis the existence of a sinusoid – corresponding to a deterministic seasonal component – at the frequency of interest, and test whether spectral density estimates warrant such a hypothesis. The early literature on spectral peak testing (see Priestley (1981)) focused on this approach, and the stable-seasonality test of Lytras et al. (2007) does as well. Any stochastic seasonal component, conceived as a nonstationary process, can include such a stable sinusoidal component without loss of generality — this is analogous to the fact that a random walk with drift can be decomposed into a linear term (its mean) plus the purely stochastic mean zero portion. Tests for stable sinusoids focus on the deterministic part of seasonality, but are not designed

* Corresponding author.
E-mail addresses: tucker.s.mcelroy@census.gov (T. McElroy), anindya@umbc.edu (A. Roy).

https://doi.org/10.1016/j.jspi.2020.06.012
0378-3758/Published by Elsevier B.V.

to address the stochastic portion. However, it is important to do so: removal of the deterministic portion alone (say, via regression) does not entail the removal of the whole stochastic seasonal, and such an approach fails to accomplish the goal of seasonal adjustment — for most economic series, the seasonality is too evolutive to be adequately captured by fixed periodic functions (Findley et al., 2017).

In summary, seasonal adjustment will typically remove deterministic seasonality as well as any nonstationary stochastic facets to seasonality. However, residual seasonality is deemed to be present if seasonal peaks exist in the spectral density of the trend-differenced adjusted process. Tests for residual seasonality can therefore be based upon a stationary time series process, and all the earlier work has recognized this. Early work on assessing the effect of seasonal adjustment appeared in Nerlove (1964) and Grether and Nerlove (1970). Pierce (1976, 1979) looked at adequacy of seasonal adjustment by examining the magnitude of the autocorrelations at seasonal lags of the adjusted series. This approach is generalized to the Q_s statistic, adopted by TRAMO-SEATS (Maravall, 2012), which is a variant of the Box–Ljung–Pierce test applied to seasonal lag autocorrelations.

Whereas a deterministic sinusoid corresponds to a jump in the spectral distribution function – and will appear in a spectral density estimate as a tall slender peak – stochastic seasonality (i.e., the residual seasonality remaining after seasonal adjustment) instead corresponds to a broader peak in the spectral density estimate (e.g., computed using an Autoregressive Estimator); it will have a broader peak that nonetheless is approaching an infinite height as sample size increases. It becomes important to consider the width of a spectral peak in its assessment. Since the procedure of seasonal adjustment, viewed in the frequency domain, amounts to multiplication of a function with a peak by another function with a trough, whether or not the peak is transformed into a trough or not depends on the width of these functions (see McElroy and Roy (2017)). For this and other reasons, Soukup and Findley (1999) considered a measure of the peak that involved the distance between ordinates of the log spectrum when examined on a grid of frequencies of mesh size $\pi/60$.

The actual measure proposed in Soukup and Findley (1999) – which has now become somewhat of a standard by virtue of its incorporation into the X-12-ARIMA software used at most international statistical agencies – is computed by comparing the spectral peak ordinate to both nearest neighbor ordinates, with respect to the chosen frequency grid; when both ordinate differences exceed a threshold (selected based upon empirical criteria), the spectral peak is declared to be "visually significant." So far no distribution theory has been proposed for this statistic, making it difficult to rigorously determine Type I and II error. This paper puts the concept of visual significance upon a rigorous statistical footing by describing the exercise of finding a visually significant peak as a proper hypothesis testing procedure.

In order to develop a statistical theory for spectral peaks, one first requires a theory for spectral density estimation. Spectral density estimates generally fall into two classes: model-based (e.g., the Autoregressive spectral estimator, or other estimators derived from a fitted model) and nonparametric (e.g., based on smoothing the periodogram). We focus on the latter class, based upon tapering the sample autocovariances with a positive definite taper, such as the Bartlett or Daniell kernels (Priestley, 1981). Asymptotic theory for such estimates goes back to Parzen (1957), and the literature adopts the perspective that the taper bandwidth is negligible relative to sample size. More recent literature, as in Hashimzade and Vogelsang (2008), adopts the perspective of the so-called fixed-b asymptotics, where the ratio of bandwidth length to sample size is assumed to be a fixed fraction $b \in (0, 1)$. Because the fixed-b asymptotic framework has several advantages – including a superior approximation of the sampling distribution (McElroy and Politis, 2014; Sun, 2014) – we pursue spectral peak detection with this perspective in mind. Some of the results require extensions of previous literature, such as Hashimzade and Vogelsang (2008) and McElroy and Politis (2014); we allow the frequencies of interest to depend on sample size, and can be more general than Fourier frequencies.

2. Spectral peaks: measurement and detection

2.1. Measuring spectral peaks

The idea of a peak in a graph is surprisingly subtle to capture through mathematical formulas. McElroy and Holan (2009) set forth a measure based upon measuring the second derivative of the spectral density — referred to as the Spectral Convexity (SC) test. The approach of Soukup and Findley (1999) compares the log autoregressive spectrum at a frequency of interest θ (in units of radians), e.g., a seasonal or trading day frequency, to two "nearest neighbors" on the left and right, some distance δ away. That is, we have three frequencies $\theta - \delta, \theta, \theta + \delta$, and comparisons are given by $\log f(\theta) - \log f(\theta - \delta)$ and $\log f(\theta) - \log f(\theta + \delta)$ for the left and right hand, respectively. Note that no peak is present if one of these differences is non-positive. But even when both differences are positive, the actual shape of the peak may be so mild as to be indistinguishable from the overall shape of the spectral graph. Clearly this also depends on the spread between the three frequencies as well.

The Visual Significance (VS) approach of Soukup and Findley (1999) is to first fix a fairly uniform grid by dividing $[0, \pi]$ into 61 frequencies. We use the qualifier "fairly," since the trading day frequencies can also be considered, and do not fall exactly on the grid points $\pi j/60$, although the seasonal frequencies corresponding to monthly or quarterly series certainly do. Thus the distance to the nearest neighbors is $\delta = 1/60 \approx 1.7\%$ of the total width. These differences – computed in terms of the log spectrum $\log f$ – must both be above a pre-specified threshold τ_f, which is taken to be the fixed fraction 6/52 of the range of the logged spectral estimate, i.e., the difference between the maximum and minimum value of the log spectrum. The fraction 6/52 was suggested by Soukup and Findley (1999) based on empirical considerations, but in our development we allow τ_f to be user-determined, allowing for additional flexibility.

A peak functional that corresponds to the Visual Significance approach of Soukup and Findley (1999) can be defined via

$$\Theta_{\theta,\delta}[g_f] = \min\{g_f(\theta) - g_f(\theta-\delta), g_f(\theta) - g_f(\theta+\delta)\} \tag{1}$$

where g_f is a suitably chosen monotone transformation of the spectral density f. Soukup and Findley (1999) chose $g_f = \log f$. We will refer to $\Theta_{\theta,\delta}[\log f]/R_{\log f}$ as the VS functional at θ (or simply VS for short), where $R_{\log f} = (\max \log f - \min \log f)$ is the range of the log spectral density (assumed to be finite). (It is intuitively appealing to let the threshold depend on the underlying spectral density, thereby making assessment of peak strength on a relative scale.) Note that $\Theta_{\theta,\delta}$ is *not* a linear functional of spectra, due to the minimum in its definition. The criterion of VS states that a visually significant peak exists at frequency θ if $\Theta_{\theta,\delta}[\log f]$ exceeds $\tau_f = \tau R_{\log f}$. In the specific settings of Soukup and Findley (1999), we would consider whether $\Theta_{j\pi/6,\pi/60}[\log f]$ (for $j = 1, 2, \ldots, 6$) exceeds τ_f, where τ is taken to be the fraction 6/52. Of course, altering this τ_f to other values will naturally change the notion of peak, e.g., lowering τ_f to 0.1 makes the VS criterion less demanding.

2.2. Peak detection

Our main goal is to develop a hypothesis testing framework for the peak detection problem. To do so we must clearly describe the null and the alternative hypotheses in terms of parameters/functionals of the spectral density. Intuitively, if there is a peak functional $P[f]$ which captures the peak strength, absence of a peak would be quantified by the event that $P[f]$ fails to exceed some threshold. Because we are particularly interested in evaluating the VS framework within the paradigm of statistical significance, we closely adhere to the concepts and notions developed by Soukup and Findley (1999). Thus, we choose $P[f] = \Theta_{\theta,\delta}[\log f]$.

The null region for the peak testing problem must include all cases which the user would generally classify as a non-peak. This includes spectral features that resemble half-peaks, where on one side of the target frequency the measure fails to qualify as peaked, even though on the other side the measure may indicate a strongly peaked feature. We denote the right-side and left-side peak measures included in the VS functional as $\Theta^L_{\theta,\delta}[\log f] = \log f(\theta) - \log f(\theta-\delta)$ and $\Theta^R_{\theta,\delta}[\log f] = \log f(\theta) - \log f(\theta+\delta)$. Thus, the quantity of interest is $\Theta_{\theta,\delta}[\log f] = \min\{\Theta^L_{\theta,\delta}[\log f], \Theta^R_{\theta,\delta}[\log f]\}$. Recognizing that an insignificant peak is one where at least one side is not significantly peaked, we obtain the following null and alternative hypotheses:

$$H^0_\theta : \Theta_{\theta,\delta}[\log f] \le \tau_f \quad \text{vs} \quad H^a_\theta : \Theta_{\theta,\delta}[\log f] > \tau_f. \tag{2}$$

Intuitively, a critical region for such a test (with Type I error approximately equal to α) should take the form

$$\mathcal{C}_\alpha = \{\widehat{f} : \min\{\Theta^L_{\theta,\delta}[\log \hat{f}], \Theta^R_{\theta,\delta}[\log \hat{f}]\} > c_\alpha + \tau_f\}, \tag{3}$$

where $\widehat{f}$ is some spectral estimator based on a time series sample of length n. Also, c_α is a constant chosen based on the joint distribution of $\Theta^L_{\theta,\delta}[\log \hat{f}]$ and $\Theta^R_{\theta,\delta}[\log \hat{f}]$, so that the size of the test (up to large sample approximation) is

$$\sup_{f:\, \Theta_{\theta,\delta}[\log f] \le \tau_f} P(\mathcal{C}_\alpha) \simeq \alpha,$$

where $\simeq$ stands for "asymptotically equal to." Because $P(\mathcal{C}_\alpha)$ is increasing in $\Theta_{\theta,\delta}[\log f]$, the supremum is achieved for f such that $\Theta_{\theta,\delta}[\log f] = \tau_f$. The test that rejects the hypothesis of no peak based on a rejection region of the form (3) will be called the VS_{new} test, in distinction to the existing Visual Significance approach that uses the VS functional as a diagnostic tool. Below we consider conditions for which $\sup_{f:\, \Theta_{\theta,\delta}[\log f] \le \tau_f} P(\mathcal{C}_\alpha) \simeq \alpha$. If the threshold τ_f is not known, a suitable sample estimate $\widehat{\tau_f}$ could be substituted.

Under the framework of Section 3 below, the distribution theory of the spectral estimator $\widehat{f}$ yields asymptotically pivotal quantities $(X_L, X_R) \equiv (\Theta^L_{\theta,\delta}[\log \hat{f}] - \Theta^L_{\theta,\delta}[\log f], \Theta^R_{\theta,\delta}[\log \hat{f}] - \Theta^R_{\theta,\delta}[\log f])$. For the peak testing problem, the marginal distributions of X_L and X_R are the same under the null hypothesis.

Proposition 1. *Let X_L and X_R be the asymptotic pivots constructed from the left and right peak functionals. Let c_α be the upper α-percentile of the asymptotic distribution of X_L (or X_R). Then the critical region (3) is an asymptotically size α test for the hypothesis (2).*

Remark 1. Proposition 1 shows that in order to bound Type I error, the marginal pivot distribution should be used; this notion is further explored via simulation in Section 4.3.

The hypothesis about the peak functional defined as the minimum of two quantities can be rewritten in terms of the bivariate measure involving the left and the right side functionals separately. Then the hypothesis testing framework for testing (2) is analogous to one sided tests for the minimum of two parameters; see Sasabuchi (1980), Berger (1989), and Liu and Berger (1995). The Type I error probability of such a test with nominal level α is typically around α^2 when both parameters are close to their null value. This feature of the test is observed in our simulation study. As observed in Sasabuchi (1980) and Berger (1989), more powerful tests can be devised using the bivariate quantity than those defined

simply using the minimum of the two random variables. This will be particularly true when the asymptotic distribution of the pivot is normal, e.g., in the case of an AR-spectral estimator. However, since the main focus of this article is to evaluate the VS functional within a hypothesis testing paradigm, we will not elaborate on bivariate tests any further.

To further develop the testing framework we need to describe the following:

1. What is the spectral estimator $\widehat{f}$?
2. What is the relevant distribution theory for $\Theta_{\theta,\delta}[\log \hat{f}]$ that would provide appropriate pivotal quantities, which can then be inverted to obtain the α-critical region $\mathcal{C}_\alpha$?
3. How should one choose τ_f (or $\widehat{\tau_f}$)?

We develop a suitable spectral estimator $\widehat{f}$ and relevant distribution theory in the next section. We end this section with a brief discussion about the choice of τ_f and its estimated value $\widehat{\tau_f}$ that is to be plugged into the formula for the critical region (3).

Choice of τ_f: For Soukup and Findley (1999), τ_f was chosen as $\tau_f = \tau R_{\log f}$ where $R_{\log f}$ is the range of the log spectrum (necessarily assumed to be bounded) and τ is a fixed constant equal to 6/52. Thus, Soukup and Findley (1999) considered a peak functional estimate not visually significant if the peak functional was less than 6/52 of the range of the estimated log-spectrum. Their choice of τ was guided by the available resolution in the plotting device at the time of development (i.e., the star plot). The choice of τ is still a subjective issue. Here we chose $\tau = 0.1$ (a value close to 6/52) based on analyses of spectral densities and their associated autocorrelations (McElroy and Roy, 2017). Simulation results are presented to demonstrate that the proposed test works for any given value of τ.

Choice of $\widehat{\tau_f}$: To develop the testing theory we assume that the true log-spectral density is bounded in absolute value (an assumption that is implicit in the definition of the VS functional). To estimate τ_f, we need to estimate $R_{\log f}$, the range of the log-spectral density. We use an estimate of f obtained using the `pspectrum` function in R. One could use other available estimators that are based on consistent estimation procedures.

3. Spectral estimation and asymptotic critical values

We consider tapered autocovariance function (acf) estimators defined as linear weighted combinations of the first M sample autocovariances, with the weights determined by a specified taper. The quantity M is the bandwidth length. Depending on how fast the bandwidth length grows relative to sample size, three different convergences are obtained. These are the cases of fixed bandwidth length, small bandwidth fraction, and fixed bandwidth fraction (fixed-b), respectively. Following recent work on fixed bandwidth fraction asymptotics (Hashimzade and Vogelsang, 2008; McElroy and Politis, 2014), which establishes that incorporating bandwidth fraction into the asymptotic theory facilitates a superior approximation to the finite-sample distribution (Sun, 2014), we concentrate on this case for the rest of the paper. Note that the spectral theory here is developed for stationary series, and hence presumes transformations and/or differencing has removed any native non-stationarity.

3.1. Fixed-b spectral estimator

Let $\langle g \rangle = (2\pi)^{-1} \int_{-\pi}^{\pi} g(\omega) d\omega$. Then the true autocovariance at lag k is $\gamma_k = \langle f c_k \rangle$, where $c_k(\theta) = \cos(\theta k)$. Supposing we have a sample $X_1, X_2, \ldots, X_n$, let us denote the periodogram via

$$\widehat{f}(\theta) = n^{-1} \left| \sum_{t=1}^{n} X_t e^{-i\theta t} \right|^2 ,$$

from which it follows that the sample acf is $\widehat{\gamma}_k = \langle \widehat{f} c_k \rangle$. The tapered acf estimator is defined as

$$\widehat{f}_{\Lambda,M}(\theta) = \sum_h \Lambda(h/M) \widehat{\gamma}_h c_h(\theta),$$

where the taper Λ is an even function that places more weight towards low lag sample acfs. Here M is the bandwidth length, chosen by the practitioner. Often the taper is compactly supported on $[-1, 1]$, so that no lags greater than M are considered in the estimator.

The fixed bandwidth fraction paradigm considers that $M/n \to b \in (0, 1]$, a fixed fraction. This b is called the bandwidth fraction and $\widehat{f}_{\Lambda,M}(\theta)$, denoted by $\hat{f}_b(\theta)$ in this case, is the fixed-b spectral estimator of f at frequency θ. The classical paradigm assumes that $M/n \to 0$, and it is known (Sun, 2014) that fixed bandwidth fraction asymptotics with small values of b (e.g., $b = .02$) essentially capture the classical asymptotics.

For developing the distribution theory for $\hat{f}_b(\theta)$ some data assumptions are needed. We present two main results below: first, the joint limiting behavior of the data's sine and cosine transforms as a functional limit theorem; second, the joint limiting behavior of the fixed-b spectral estimates at different frequencies. Excluding the long memory cases of McElroy and Politis (2014) for simplicity, we focus on three possible data generating process:

- *Process* P1. $\{X_t\}$ is a linear process: $X_t = \sum_j \psi_j \epsilon_{t-j}$ with $\{\psi_j\}$ square summable and $\{\epsilon_t\}$ iid with finite fourth moment.
- *Process* P2. $X_t = g(Z_t)$ for each t, where g is a function in $\mathbb{L}^2(\mathbb{R}, e^{-x^2/2})$, and $\{Z_t\}$ is a Gaussian process.
- *Process* P3. $\{X_t\}$ is a strictly stationary process whose kth order cumulants exist and are summable over its k indices, for all $k \geq 1$.

The assumptions for process P1 are standard, whereas the framework of P2 allows us to consider some non-linear processes. Processes P1 and P3 allow for non-Gaussian data, but P3 only allows for short-range dependence. These are viewed as unverifiable conditions that instead indicate the range of processes for which we can expect validity.

We are interested in establishing joint Discrete Fourier Transform (DFT) results over a sequence of frequencies near a finite set of target frequencies. Let the target frequencies (for J fixed) be $\{\theta_1^0, \ldots, \theta_J^0\}$, which allows a general formulation; however, in our application the target frequencies are all the same, i.e., $\theta_j^0 = \theta^0$ for all j. Let $\{\theta_{1,n}, \ldots, \theta_{J,n}\}$ be a sequence of frequencies over which we will evaluate the large sample properties of the DFTs. For any typical frequency in the sequence and the corresponding target, we use the notation $\overline{\theta}_j$ to denote the leading n^{-1} term in the frequency, i.e.,

$$\theta_{j,n} = \theta_j^0 + n^{-1}\,\overline{\theta}_j + o(n^{-1}). \tag{4}$$

In particular, $\overline{\theta}_j = \lim_{n\to\infty} n\,(\theta_{j,n} - \theta_j^0)$ and does not depend on n. Thus, the θ_j^0 are fixed target frequencies being approached at a certain rate over the sequence of frequencies $\theta_{j,n}$.

The real and the imaginary parts of the DFT at some frequency θ are the cosine and the sine transforms at that frequency, and they are defined as

$$S_n^c(\theta) = \sum_{t=1}^n X_t \cos(\theta t) \qquad S_n^s(\theta) = \sum_{t=1}^n X_t \sin(\theta t). \tag{5}$$

The distribution theory used in the proposed testing procedure for the peak functional will depend on the asymptotic distribution of the sine and cosine transforms at relevant frequencies. When θ_j^0 is equal to zero or π, the asymptotic results are a bit different from the other cases where $\theta_j^0 \in (0, \pi)$. We first give an expression for the limiting covariances

$$V_{cc}(\theta_{j,n}, \theta_{k,n}) = \mathrm{cov}(S_n^c(\theta_{j,n}), S_n^c(\theta_{k,n})) \tag{6}$$
$$V_{cs}(\theta_{j,n}, \theta_{k,n}) = \mathrm{cov}(S_n^c(\theta_{j,n}), S_n^s(\theta_{k,n})) \tag{7}$$
$$V_{ss}(\theta_{j,n}, \theta_{k,n}) = \mathrm{cov}(S_n^s(\theta_{j,n}), S_n^s(\theta_{k,n})) \tag{8}$$

for a typical pair of frequencies $(\theta_{j,n}, \theta_{k,n})$. Also, for pairs of frequencies along with corresponding target frequencies, define the sets

$$\mathcal{A} = \{(\theta_j, \theta_k) \in [0,\pi] \times [0,\pi] : \theta_j^0 + \theta_k^0 = 0, 2\pi \text{ (modulo } 2\pi)\}$$
$$\mathcal{B} = \{(\theta_j, \theta_k) \in [0,\pi] \times [0,\pi] : \theta_j^0 - \theta_k^0 = 0, 2\pi \text{ (modulo } 2\pi)\}. \tag{9}$$

Proposition 2. *Let $(S_n^c(\theta_{j,n}), S_n^s(\theta_{k,n}))$ be the cosine and sine transformations (5) for frequencies $\theta_{j,n}$ and $\theta_{k,n}$ respectively, and let $V_{cc}(\theta_{j,n}, \theta_{k,n})$, $V_{cs}(\theta_{j,n}, \theta_{k,n})$, and $V_{ss}(\theta_{j,n}, \theta_{k,n})$ denote the covariances as defined in (6), (7), and (8). Then*

$$V_{cc}(\theta_{j,n}, \theta_{k,n}) = n\,\underline{V}_{cc}(\overline{\theta}_j, \overline{\theta}_k) + o(n),$$
$$V_{cs}(\theta_{j,n}, \theta_{k,n}) = n\,\underline{V}_{cs}(\overline{\theta}_j, \overline{\theta}_k) + o(n),$$
$$V_{ss}(\theta_{j,n}, \theta_{k,n}) = n\,\underline{V}_{ss}(\overline{\theta}_j, \overline{\theta}_k) + o(n),$$

where

$$\begin{aligned}\underline{V}_{cc}(\overline{\theta}_j, \overline{\theta}_k) = \frac{1}{2}&\left(\frac{f(\theta_j^0)+f(\theta_k^0)}{2}\right)\left(\frac{\sin(\overline{\theta}_j+\overline{\theta}_k)}{\overline{\theta}_j+\overline{\theta}_k}\,1_{\mathcal{A}} + \frac{\sin(\overline{\theta}_j-\overline{\theta}_k)}{\overline{\theta}_j-\overline{\theta}_k}\,1_{\mathcal{B}}\right)\\ &-(g(\theta_j^0)+g(\theta_k^0))\left(\frac{1-\cos(\overline{\theta}_j+\overline{\theta}_k)}{\overline{\theta}_j+\overline{\theta}_k}\,1_{\mathcal{A}}\right)\\ &-(g(\theta_j^0)-g(\theta_k^0))\left(\frac{1-\cos(\overline{\theta}_j-\overline{\theta}_k)}{\overline{\theta}_j-\overline{\theta}_k}\,1_{\mathcal{B}}\right),\end{aligned}$$

$$\begin{aligned}\underline{V}_{cs}(\overline{\theta}_j, \overline{\theta}_k) = \frac{1}{2}&\left(\frac{f(\theta_j^0)+f(\theta_k^0)}{2}\right)\left(\frac{1-\cos(\overline{\theta}_j+\overline{\theta}_k)}{\overline{\theta}_j+\overline{\theta}_k}\,1_{\mathcal{A}} - \frac{1-\cos(\overline{\theta}_j-\overline{\theta}_k)}{\overline{\theta}_j-\overline{\theta}_k}\,1_{\mathcal{B}}\right)\\ &+(g(\theta_j^0)+g(\theta_k^0))\left(\frac{\sin(\overline{\theta}_j+\overline{\theta}_k)}{\overline{\theta}_j+\overline{\theta}_k}\,1_{\mathcal{A}}\right)\end{aligned}$$

$$- (g(\theta_j^0) - g(\theta_k^0)) \left(\frac{\sin(\overline{\theta}_j - \overline{\theta}_k)}{\overline{\theta}_j - \overline{\theta}_k} 1_{\mathcal{B}} \right),$$

$$\underline{V}_{ss}(\overline{\theta}_j, \overline{\theta}_k) = \frac{1}{2} \left(\frac{f(\theta_j^0) + f(\theta_k^0)}{2} \right) \left(\frac{\sin(\overline{\theta}_j + \overline{\theta}_k)}{\overline{\theta}_j + \overline{\theta}_k} 1_{\mathcal{A}} - \frac{\sin(\overline{\theta}_j - \overline{\theta}_k)}{\overline{\theta}_j - \overline{\theta}_k} 1_{\mathcal{B}} \right)$$
$$+ (g(\theta_j^0) + g(\theta_k^0)) \left(\frac{1 - \cos(\overline{\theta}_j + \overline{\theta}_k)}{\overline{\theta}_j + \overline{\theta}_k} 1_{\mathcal{A}} \right)$$
$$- (g(\theta_j^0) - g(\theta_k^0)) \left(\frac{1 - \cos(\overline{\theta}_j - \overline{\theta}_k)}{\overline{\theta}_j - \overline{\theta}_k} 1_{\mathcal{B}} \right),$$

$g(\theta_j) = \sum_{h=1}^{\infty} \gamma_h \sin(\theta_j h)$, *and the sets* $\mathcal{A}$ *and* $\mathcal{B}$ *are defined in* (9).

Proposition 2 allows us to write a joint functional limit theorem for the sine and cosine transformations at the frequencies $\{\theta_{1,n}, \ldots, \theta_{J,n}\}$. In the following, Theorem 1 gives the asymptotic distribution of the sine and cosine transforms at the given sequence of frequencies. Define stochastic processes from the sine and cosine transforms via

$$S^s_{\lfloor rn \rfloor}(\theta) = \sum_{t=1}^{\lfloor rn \rfloor} X_t \sin(\theta t), \qquad S^c_{\lfloor rn \rfloor}(\theta) = \sum_{t=1}^{\lfloor rn \rfloor} X_t \cos(\theta t) \tag{10}$$

as a function of $r \in (0, 1)$, and write $\xi^s_{\lfloor \cdot n \rfloor}$ and $\xi^c_{\lfloor \cdot n \rfloor}$ for the linearly interpolated versions; see McElroy and Politis (2014, p.214). Our results below do not consider mean-centering, because in applications the time series have already been regression-adjusted and differenced to remove nonstationarity, so that typically the mean is zero. Extensions of the theorems below to a non-zero mean are possible, as outlined in McElroy and Politis (2014), but there is no impact except at frequency zero anyways, and we have no interest in peak detection at frequency zero for purposes of assessing adequacy of seasonal adjustment. For this reason, we restrict the target frequencies to $(0, \pi)$.

Theorem 1. *Let* $\{X_t\}$ *be a mean zero covariance stationary time series corresponding to one of the frameworks P1, P2, or P3. Let* $\{\theta_{1,n}, \ldots, \theta_{J,n}\}$ *be a sequence of frequencies as defined in* (4) *with target frequencies* $(\theta_1^0, \ldots, \theta_J^0)$ *all in* $(0, \pi)$. *Suppose* $\xi^x_{\lfloor rn \rfloor}$ *be the linearly interpolated versions of the partial sum processes* (10) *for* $x = c, s$. *Then as* $n \to \infty$,

$$\frac{1}{\sqrt{n}} \begin{pmatrix} \xi^c_{\lfloor rn \rfloor}(\theta_{1,n}) \\ \xi^s_{\lfloor rn \rfloor}(\theta_{1,n}) \\ \vdots \\ \xi^c_{\lfloor rn \rfloor}(\theta_{J,n}) \\ \xi^s_{\lfloor rn \rfloor}(\theta_{J,n}) \end{pmatrix} \Longrightarrow \underline{V}^{1/2} \begin{pmatrix} B^c_1(r) \\ B^s_1(r) \\ \vdots \\ B^c_J(r) \\ B^s_J(r) \end{pmatrix} := \begin{pmatrix} A^c_1(r) \\ A^s_1(r) \\ \vdots \\ A^c_J(r) \\ A^s_J(r) \end{pmatrix},$$

where each process B^c_j *and* B^s_j *for* $1 \le j \le J$ *is an independent Brownian motion. The covariance matrix of the vector* $A(r)$ *process for any fixed* r *is* $\underline{V}$, *with the covariance of* $A^x_j(r)$ *and* $A^y_k(r)$ *given by* $\underline{V}_{xy}(\overline{\theta}_j, \overline{\theta}_k)$ *in Proposition 2 for* x, y *denoting either* c *or* s.

Using the result of Theorem 1, it is possible to write down the joint distribution of the fixed-b spectral estimator $\hat{f}_b$ at finitely many frequencies. For the next result, we allow Λ to be flat-top (i.e., there is a $c \in [0, 1)$ such that $\Lambda(x)$ is constant for $|x| \le c$) and is piecewise twice continuously differentiable.

Theorem 2. *Let* $\{X_t\}$ *be a mean zero covariance stationary time series corresponding to one of the frameworks condition P1, P2, or P3. Then for the sequence of frequencies* $(\theta_{1,n}, \ldots, \theta_{J,n})$, *as* $n \to \infty$

$$\left(\frac{\widehat{f}_b(\theta_{1,n})}{f(\theta_1^0)}, \ldots, \frac{\widehat{f}_b(\theta_{J,n})}{f(\theta_J^0)} \right) \Longrightarrow (S_{\theta_1^0}(b), \ldots, S_{\theta_J^0}(b)).$$

The quantities in limiting random vector, $S_{\theta_j^0}(b)$, *are defined as*

$$S_{\theta_j^0}(b) = -b^{-2} \int\int_{cb < |r-s| < b} \ddot{\Lambda}((r-s)\, b^{-1}) (A^c_j(r) A^c_j(s) + A^s_j(r) A^s_j(s))\, dr ds$$
$$+ 2b^{-1}\, \dot{\Lambda}_-(1) \int_0^{1-b} (A^c_j(r) A^c_j(r+b) + A^s_j(r) A^s_j(r+b))\, dr$$
$$- 2b^{-1}\, \dot{\Lambda}_+(c) \int_0^{1-bc} (A^c_j(r) A^c_j(r+bc) + A^s_j(r) A^s_j(r+bc))\, dr$$

$$+ 2b^{-1} \int_{1-b}^{1-bc} \dot{\Lambda}((1-r)\,b^{-1})\,(A_j^c(r)A_j^c(1) + A_j^s(r)A_j^s(1))\,dr$$
$$+ \Lambda(0)\left(A_j^c(1)A_j^c(1) + A_j^s(1)A_j^s(1)\right),$$

where the processes A_j^c and A_j^s are as defined in Theorem 1, and $\dot{\Lambda}(x)$, $\ddot{\Lambda}(x)$, $\dot{\Lambda}_-(x)$ and $\dot{\Lambda}_+(x)$ are the first, second, left and right derivatives of Λ at x, respectively. In the case that there is a jump discontinuity in Λ at c, we must replace the third summand in the limit distribution by

$$2(\Lambda^+(c) - \Lambda^-(c))\,(A_j^c(1-bc)A_j^c(1) + A_j^s(1-bc)A_j^s(1)).$$

Allowing for jump discontinuities at c means the results cover the truncation taper, which may be important for application of the theory to other forms of estimators.

3.2. Critical value and rejection region for the VS test

For the specific application of peak testing, the finitely many frequencies $\theta_{1,n}, \ldots, \theta_{J,n}$ involved in the construction of any local measure (by local we mean that only features in the neighborhood of the target seasonal frequency are considered for the determination of a peak) of peak strength, e.g. (1), will have the property $\theta_{j,n} = \theta^0 + O(n^{-1})$. Thus, for any two frequencies $\theta_{j,n}$, $\theta_{k,n}$, the associated variance covariance expressions for the sine and cosine transforms reduce to

$$\begin{aligned}
\underline{V}_{cc}(\overline{\theta}_j, \overline{\theta}_j) = \underline{V}_{ss}(\overline{\theta}_j, \overline{\theta}_j) &= \frac{1}{2} f(\theta^0), \\
\underline{V}_{sc}(\overline{\theta}_j, \overline{\theta}_j) &= 0, \\
\underline{V}_{cc}(\overline{\theta}_j, \overline{\theta}_k) = \underline{V}_{ss}(\overline{\theta}_j, \overline{\theta}_k) &= \frac{1}{2} f(\theta^0)\,\frac{\sin(\overline{\theta}_j - \overline{\theta}_k)}{\overline{\theta}_j - \overline{\theta}_k}, \\
\underline{V}_{sc}(\overline{\theta}_k, \overline{\theta}_j) = -\underline{V}_{cs}(\overline{\theta}_k, \overline{\theta}_j) &= \frac{1}{2} f(\theta^0)\,\frac{1 - \cos(\overline{\theta}_j - \overline{\theta}_k)}{\overline{\theta}_j - \overline{\theta}_k}, \\
\underline{V}_{cc}(\overline{\theta}_k, \overline{\theta}_k) = \underline{V}_{cc}(\overline{\theta}_k, \overline{\theta}_k) &= \frac{1}{2} f(\theta^0), \\
\underline{V}_{sc}(\overline{\theta}_k, \overline{\theta}_k) &= 0.
\end{aligned} \tag{11}$$

The critical value of the proposed test will be obtained from the distribution of $S_\theta(b)$ as given in Theorem 2 with the elements of the covariance matrix of the $A(r)$ process given by (11). If

$$\Theta_{\theta,\delta}[\log S(b)] = \min\left\{\log \frac{S_\theta(b)}{S_{\theta-\delta}(b)}, \log \frac{S_\theta(b)}{S_{\theta+\delta}(b)}\right\},$$

then the critical value, c_α, used in the $\Theta_{\theta,\delta}[\log f]$ test for a nominal level α is the upper $100(1-\alpha)\%$ point of the distribution of $\Theta_{\theta,\delta}[\log S(b)]$. The rejection region for the proposed VS_{new} test will be given by

$$\widehat{\mathcal{C}}_\alpha = \{\hat{f}_b : \Theta_{\theta,\delta}[\log \hat{f}_b] > c_\alpha + \tau\, R_{\log \hat{f}}\},$$

where $R_{\log \hat{f}}$ is the estimated range of log-spectrum based on a consistent estimator $\hat{f}$ of f and $\hat{f}_b$ is the fixed-b estimator of f.

4. Simulation studies

4.1. AR(2) peak

We start evaluating the peak detection performance of the proposed VS test using an AR(2) process $\{X_t\}$ satisfying

$$(1 - 2\rho\,\cos(\theta)\,B + \rho\,B^2)X_t = \epsilon_t \tag{12}$$

with noise variance $\sigma_\epsilon^2 = 1$. The parameterization in (12) puts a single peak at the frequency $\theta = \pi/6$. The values of ρ that makes the VS values at $\omega = \pi/6$ to be $\{0, 0.05, 0.10, 0.15, 0.20, 0.30, 0.40, 0.50\}$ for a width of $\pi/30$ are $\rho = \{0.000, 0.861, 0.915, 0.941, 0.959, 0.980, 0.9918, 0.9976\}$. Fig. 1 shows the spectrum for the different VS values.

Table 1 shows the power (based on a nominal rate $\alpha = .05$) for testing for the peak at a given location, i.e., $\omega = \pi/6$. The best results were obtained for bandwidth, b, in the neighborhood of 0.5 and only those are reported. Smaller values of b (e.g. 0.05) provided poorer results (very conservative with low power) compared to $b = 0.5$ and are not reported here. Since the smaller value corresponds to those obtained from the classical approach of bandwidth approaching zero (Sun, 2014), the results provide a justification for using the fixed-b approach over classical approach. The first three rows correspond to null levels of τ, and hence the rejection rates in these rows are Type I error probabilities. The proposed

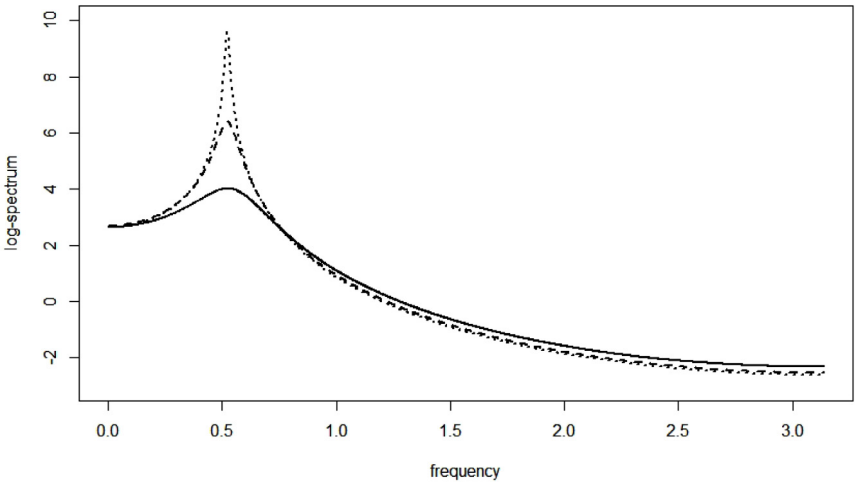

Fig. 1. AR(2) spectral densities with peak at seasonal frequency $\pi/6$. The densities with the lowest to the highest peak at $\pi/6$ correspond to VS values 0.1 (solid), 0.3 (dot-dashed), and 0.5 (dashed), respectively.

Table 1
Size and power of proposed tests based on different values of VS in the AR(2) process (12), where VS ≤ 0.1 are null values. The column *SC* shows the power of the spectral convexity test (SC). The sample sizes are $n = 120$ and $n = 600$, and the width δ used in the VS measure is $\pi/30$.

VS	$n = 120$				$n = 600$			
	SC	Bandwidth fraction b			SC	Bandwidth fraction b		
		0.4	0.5	0.6		0.4	0.5	0.6
0.00	0.010	0.002	0.006	0.009	0.063	0.018	0.017	0.020
0.05	0.013	0.001	0.005	0.008	0.625	0.003	0.004	0.007
0.10	0.029	0.001	0.004	0.006	0.893	0.026	0.025	0.025
0.15	0.045	0.002	0.019	0.027	0.905	0.078	0.073	0.070
0.20	0.042	0.011	0.050	0.065	0.912	0.233	0.205	0.202
0.30	0.024	0.031	0.117	0.158	0.929	0.787	0.731	0.703
0.40	0.009	0.095	0.367	0.418	0.841	0.990	0.981	0.975
0.50	0.003	0.351	0.732	0.757	0.359	0.999	0.997	0.996

test maintains the nominal level. For the smaller sample size, the test is very conservative but reaches the nominal size at $n = 600$.

For comparison we also provide the rejection probability of the spectral convexity (SC) test from McElroy and Holan (2009). The SC test is oversized for larger sample sizes; however, the SC test was designed for a null value of $\tau = 0$. Thus, the rejection rate is appropriate for the VS $= 0$ case. The test is not flexible enough to adapt to other null values of τ; also, the power is maximized for values of VS around .2, which is due to the convexity being positive over a wide band as the peak gets more narrow — see Fig. 1.

4.2. *AR*(14) *peak*

We consider a family of AR(14) models with a peak location fixed at a particular frequency (chosen to be $\pi/6$ without loss of generality). An AR(14) is flexible enough to capture many features of a spectral peak such as additional troughs and crests, asymmetry, shoulders, etc. In general, the difficulty in using such a model for evaluating performance of the test at a single frequency is that the VS at that frequency cannot be varied independent of other features in the model. A more careful parameterization can generate classes of complex AR(14) spectral densities, where the peak measure at a given location is varied essentially independently of the spectral features at other frequencies. To that end, consider an AR(14) model of the form

$$\prod_{i=1}^{7}(1 - 2\rho_i \cos(\theta_i)B + \rho_i^2 B^2)X_t = \epsilon_t, \tag{13}$$

where the parameters $(\rho_2, \ldots, \rho_7)$ and $(\theta_2, \ldots, \theta_7)$ are fixed throughout the simulation experiment. The parameters (ρ_1, θ_1) are varied to generate spectral densities with a peak at $\pi/6$ with specified VS values. To generate VS values

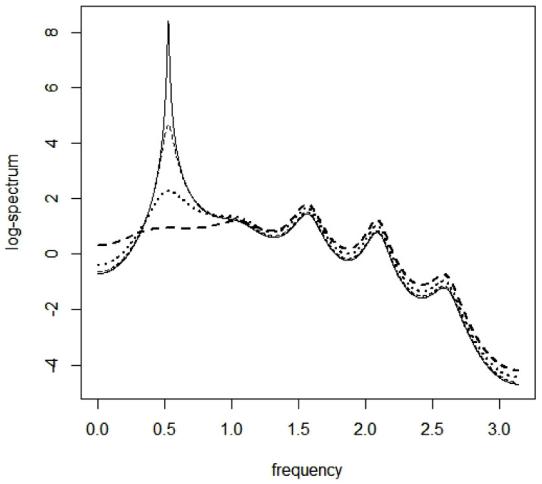

Fig. 2. AR(14) spectral densities with peak at seasonal frequency $\pi/6$. The densities with the lowest to the highest peak at $\pi/6$ correspond to VS values 0.05 (dashed), 0.1 (dotted), 0.3 (dot-dashed), and 0.5 (solid), respectively.

Table 2
Size and power of proposed tests based on different values of VS in the AR(14) process (13), where VS ≤ 0.1 are null values. The column *SC* shows the power of the spectral convexity test (SC). The sample sizes are $n = 120$ and $n = 600$, and the width δ used in the VS measure is $\pi/30$.

VS	$n = 120$				$n = 600$			
	SC	Bandwidth fraction b			SC	Bandwidth fraction b		
		0.4	0.5	0.6		0.4	0.5	0.6
0.000	0.084	0.000	0.000	0.003	0.660	0.000	0.001	0.003
0.050	0.181	0.001	0.003	0.009	0.736	0.010	0.010	0.011
0.100	0.225	0.001	0.009	0.014	0.767	0.024	0.025	0.024
0.150	0.225	0.018	0.032	0.048	0.740	0.114	0.109	0.106
0.200	0.255	0.045	0.075	0.098	0.793	0.294	0.261	0.255
0.300	0.232	0.098	0.219	0.260	0.808	0.823	0.785	0.758
0.400	0.143	0.233	0.500	0.552	0.772	0.984	0.978	0.974
0.500	0.080	0.658	0.804	0.823	0.685	1.000	0.999	0.998

0, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4 and 0.5 when the VS measure is defined with a width of $\pi/30$, the associated (ρ_1, θ_1) values used are

$$\rho_1 \in \{.99435, .9831, .9627, .929, .9042, .8715, .822, .734\}$$
$$\theta_1 \in \{.00000, .0005, .0030, .013, .0250, .0450, .099, .342\}.$$

The rest of the parameters were held fixed at

$$(\rho_2, \ldots, \rho_7) = (.67, .13, .8, .9, .92, .9)$$
$$(\theta_2, \ldots, \theta_7) = (\pi/6, 2\pi/6, 2\pi/6, 3\pi/6, 4\pi/6, 5\pi/6)$$

and the noise variance $\sigma_\epsilon^2 = 1$. The spectral densities for different VS values at $\pi/6$ are shown in Fig. 2. One can see that the peak features are co-mingled with features from neighboring peaks, particularly when the VS value is low.

We applied the proposed VS_{new} testing procedure on simulated AR(14) Gaussian processes. Table 2 shows the power for testing for the peak at a given location, e.g., $\omega = \pi/6$. The best results were obtained for bandwidth, b, in the neighborhood of 0.5 and only those are reported. The proposed test maintains the nominal level. For the smaller sample size, the test is conservative but reaches the nominal size at $n = 600$. The SC procedure continues to be over-sized.

4.3. Spectral peaks with multiple features

Ideally, to understand the contribution of any single attribute one could observe the change in the peak resulting from changing the value of that attribute while fixing the remaining attributes. Such an exercise is not possible in the AR models, where the attributes are inter-related through the parameterization of the model. As a consequence, for AR models there is no obvious peak functional which combines the attributes into a single measure and behaves monotonically with respect to any qualitative change in the peak. However, it is possible to define a new class of spectral models that parameterizes

Table 3
Size and power of proposed tests based on different values of VS in the parametric process (A.1), where VS ≤ 0.1 are null values. The column *SC* shows the power of the spectral convexity test (SC). The sample sizes are $n = 120$ and $n = 600$, and the width δ used in the VS measure is $\pi/30$.

VS	$n = 120$				$n = 600$			
	SC	Bandwidth fraction b			SC	Bandwidth fraction b		
		0.4	0.5	0.6		0.4	0.5	0.6
0.000	0.003	0.000	0.000	0.000	0.044	0.000	0.000	0.002
0.050	0.011	0.000	0.004	0.005	0.612	0.006	0.006	0.006
0.100	0.050	0.001	0.011	0.018	0.952	0.016	0.019	0.019
0.150	0.120	0.005	0.018	0.025	0.962	0.046	0.044	0.044
0.200	0.155	0.020	0.034	0.042	0.972	0.064	0.056	0.052
0.300	0.194	0.042	0.062	0.092	0.964	0.192	0.168	0.148
0.400	0.195	0.097	0.136	0.156	0.972	0.418	0.360	0.346
0.500	0.125	0.158	0.248	0.266	0.964	0.652	0.592	0.542

Table 4
Type I error (left panel) of the VS_{new} test for model (A.1) for different values of the right VS measure when the left VS is fixed at the null value $\tau = 0.1$. The sample size is $n = 600$ and the width δ used in the VS measure is $\pi/15$. For comparison, the rejection rates of the existing VS rule are given in the right panel.

VS	VS_{new}			VS rule		
	Bandwidth fraction b			Bandwidth fraction b		
	0.4	0.5	0.6	0.4	0.5	0.6
0.10	0.021	0.021	0.022	0.256	0.261	0.265
0.15	0.026	0.025	0.026	0.301	0.306	0.308
0.20	0.035	0.036	0.037	0.341	0.346	0.349
0.30	0.052	0.052	0.054	0.390	0.395	0.399
0.40	0.056	0.056	0.058	0.399	0.409	0.418
0.50	0.057	0.058	0.060	0.396	0.406	0.416

the relevant attributes separately, such that a peak functional of the form (1) will accurately quantify the notion of a peak; Appendix A of the supplementary material provides further details.

For simulation we first chose symmetric peaks where the width, the ratio of the peak height to the base height and the convexity were identical for the left and the right parts of the peaks. We chose the width of the peak to be $\pi/12$, making it wider than the width δ used in the definition of VS. This choice would ensure that VS is not simply based upon the ratio of the peak height and the base height, which we chose to be 100. The convexity parameter was obtained by fixing the width, the ratio and the VS value. Table 3 provides the size and the power of the VS_{new} test for the single peak model (A.1) for sample sizes $n = 120, 600$. Given that the peak is symmetric, the Type I error rate at the boundary (VS $= 0.1$) is expected to be around $\alpha^2 = 0.025$ (see the discussion after Proposition 1), and that is corroborated by the results. As expected, the power grows at a faster rate for the larger sample size when the VS value moves away from the null. By comparison, the SC procedure is over-sized.

4.4. Size of the VS_{new} *test and the VS rule*

For the piecewise model (A.1), it is possible to look at the performance of the test when the peak is asymmetric and investigate the Type I error rate for null values where one side of the VS has a high value while the other side is within the null region. We set the left side VS value to be equal to the null boundary value of $\tau = 0.1$, and vary the right side VS value between 0.1 and 0.5. As expected (cf. Proposition 1), the Type I error rate approaches the nominal error rate α when the right VS becomes large, and hence the peak at the right side becomes significantly large relative to the dynamic range; see Table 4. The test has a Type I error slightly larger than α for large right VS value, but that is due to finite sample size.

The existing Visual Significance approach uses a rule that declares the presence of a spectral peak if the VS functional value exceeds the 6/52 fraction of the observed dynamic range. A comparison of this rule – which we henceforth refer to as the "VS rule" – in the proposed testing framework would amount to a rejection rule (3) where the percentile c_α is set to zero. This would lead to inflated Type I error. From Table 4 (right panel) we see that the VS rule is indeed severely over-sized. To make the results comparable with the proposed testing method, we have used the fraction $\tau = 0.1$ as the threshold for the VS rule, but the Type I error rate at 6/52 is inflated as well.

There are other seasonality tests, such as the Q_s statistic of (X-13ARIMA-SEATS 2015; Maravall (2012)) and the stable-seasonality test (Lytras et al., 2007), that are available in the literature, but they are not appropriate for comparison. The Q_s test does not have a rigorously established distribution theory (Findley et al., 2017) and can yield false detections of seasonality in non-seasonal processes. The stable-seasonality test is developed in time domain for deterministic cyclical components, and hence is not designed for stationary stochastic seasonality.

5. Data analysis

The X-13ARIMA-SEATS software allows users to apply VS to the raw (original) data, to the RegARIMA model residuals, the seasonally adjusted (SA) series, or the estimated irregular. However, the raw data typically has nonstationary features, so that stationary tapered spectral estimators are not appropriate, and the methods of this paper should not be applied. More precisely, the tools of this paper properly apply to stationary processes, and cannot be applied to nonstationary processes without a huge distortion to the asymptotic distribution theory. Whereas raw data will typically exhibit trend and seasonal nonstationary features, SA data can be assumed to have only milder forms of seasonality present; trend-differencing the SA series often makes it appear more like data arising from a stationary process. Our methods can then be applied to trend-differenced SA data, or to model residuals, or the estimated irregular.

Since the critical values of our procedure are based on a stationary null hypothesis, its application to raw data may merely result in a very high rejection rate. Because the distributional properties for the nonstationary case are unknown, we do not make this application. Both the model residuals and irregular are supposed to exhibit an absence of nonstationary features, and residual spectral peaks indicate potential problems with the adjustment.

5.1. Empirical analysis of series from multiple sectors

For data analysis, we consider monthly series from four different sectors: manufacturing (86 series), retail (6 series), wholesale (18 series) and residential construction (16 series). Each series begins in April 1995, and is 20 years long. We processed each series in X-13ARIMA-SEATS and noted the visually significant peaks detected by the VS diagnostics for the seasonally adjusted series. In order to reduce the impact of human decisions on the adjustments, automatic modeling and outlier detection was used by X-13ARIMA-SEATS, along with trading day identification. We applied the VS_{new} test with the Bartlett spectral estimator to differenced seasonal adjustments, model residuals, and irregulars, reporting the p-values in each case and utilizing a null τ value of 0.1. This test was based upon the last 10 years of data to make it comparable with the simulation setting; the default VS procedure uses the final 8 years of data.

In terms of results, all but one of the adjustments were adequate when the X-13ARIMA-SEATS VS criterion is used. The only series that the VS criterion found to have residual seasonality after adjustment was the manufacturing series (series 40: transportation equipment). The VS criterion indicated that there is a significant peak in the fourth seasonal frequency. Our test also found inadequate adjustment in one manufacturing series (series 22: material handling equipment) at the third seasonal frequency. The rest of the series were deemed to have an adequate adjustment. A closer look at the spectrum (based on Bartlett spectral estimator) of the seasonally adjusted series for manufacturing series 22 and 40 (Fig. 3) reveals a peak-like feature at the third seasonal frequency for series 22, where the proposed test flagged a significant peak; however, for series 40 there is actually a trough-like feature at the fourth seasonal frequency, indicating that VS criterion finding is most likely a false positive.

5.2. Diagnosis of residual seasonality after inadequate adjustment

We also examined 15 monthly retail series with suspected seasonal behavior, and intentionally performed inadequate seasonal adjustment to check if the test could detect residual seasonality after the adjustment. The series titles are given in Table 5, with start and end dates of January 1992 and December 2007, respectively. Thus the length of each series was $n = 192$.

The series are suspected to have changing seasonality (based on the fact that X-13ARIMA-SEATS chose short filters for each of the series) and hence each calendar month's average seasonality for the last 8 years differs from that for the 16 year average. Thus, we expect to see residual seasonality in the last 8 years of data for most of the series after each series has been adjusted for seasonality using 16 year monthly averages, i.e., after replacing each value by its difference from the mean for that month over 16 years. Fig. 4 shows the means of first differences of logarithmic data by calendar months for the 15 series, where the means are computed based on the last 8 years of data. The plots are all on the same scale. The plots show a varying degree of month to month changes in the mean plots, substantiating the claim of residual seasonality for most of the series.

We applied the VS_{new} test along with the existing VS diagnostic procedure in X-13ARIMA-SEATS to the last 8 years of data for each of the series. The VS procedure from X-13ARIMA-SEATS was applied to the spectrum estimated both using an AR(30) model as well as via the raw periodogram. The spectral diagnostic tests based on the AR(30) spectrum and periodogram as well as the VS_{new} test are applied to each of the five seasonal frequencies individually. The tests for individual seasonal frequencies were not adjusted for multiple testing.

Table 5 reports the peak detection results for all the tests for the 15 series considered. The columns arspec s-pk and pdg s-pk correspond to the X-13ARIMA-SEATS VS diagnostics computed based on the AR(30) approximation of the spectrum and the raw periodogram, respectively. For the spectrum diagnostics and the proposed VS_{new} test the columns list the seasonal frequencies with significant seasonality.

Except for the VS criterion based on the raw periodogram, the diagnostics/tests detect seasonality in most of the series. The total number of series where some seasonality was detected is given at the bottom row. Overall, the procedures have comparable performance. The VS criterion based on autoregressive spectral estimate detects more peaks than the rest of

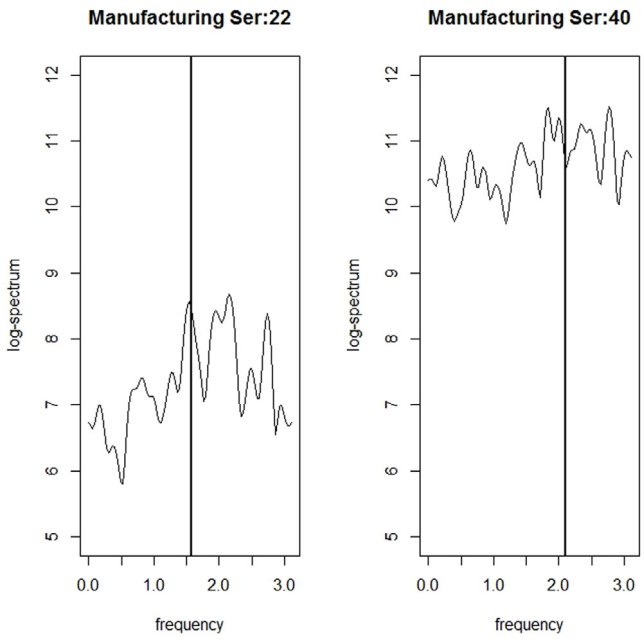

Fig. 3. Spectral plot for seasonally adjusted series for manufacturing series 22 and series 40. The vertical lines indicate the third seasonal (left panel) and the fourth seasonal (right panel) frequencies.

Table 5
Diagnostic indications of residual seasonality in the last 10 Years of the stable-seasonal adjustment, for 15 retail series delineated in the first column. The second column pertains to the AR(30) VS method of X-13ARIMA-SEATS, whereas the third column utilizes the periodogram VS method.

Title	arspec s-pk	pdg s-pk	VS_{new}
Retail and food services sales, total	1,2,3	1,3	2
Electronics and appliance stores	2,4 –	–	2,4
Computer and software stores	3	–	4
Building materials and supplies dealers	2,5	–	2
Grocery stores	1	–	–
Clothing and clothing accessory stores	1,3	–	1,3,4
Women's clothing stores	1,3	–	1,2
Shoe stores	4	2	1
Sporting goods, hobby, book, and music stores	–	–	2
General merchandise stores	1,2,3	1,2	1,2,4
Department stores – excluding leased departments	2	1	1,2
Warehouse clubs and superstores	1,2,5	2	1,2,5
Nonstore retailers	1,2	1	2
Electronic shopping and mail-order houses	1,2	1	1,2
Food services and drinking places	1	1	–
Totals	15	8	13

the tests. The VS_{new} test usually detects a subset of VS peaks as statistically significant peaks with a few discrepancies. The VS_{new} test was insignificant for two series: "Grocery stores" and "Food services and drinking places". A closer look reveals the reason why the VS_{new} test did not show any significant peaks for these series. For the "Grocery stores" series and the "Food services and drinking places" the monthly mean plot in Fig. 4 shows that the magnitude of month to month movement in the last 8 years of the adjusted series are minimal, indicating a low degree of moving seasonality in these series. Also, the spectral estimates for the series do not show any significant peak feature for these two series (last plots in rows 1 and 3 of Fig. 5). Moreover, the p-values for the test statistics for the proposed test were all more than 0.5 for all of the seasonal frequencies, except the third and fourth seasonal frequencies for the grocery store series where the p-values were 0.09 and 0.1, respectively. Overall, the findings correspond to the features exhibited in the monthly mean plot and the spectral estimate plot.

6. Discussion

Because seasonal adjustment is an enormous activity for statistical offices, the determination of adjustment inadequacy is extremely important. A host of criteria have been proposed over the decades (summarized in Hylleberg (1986)), and

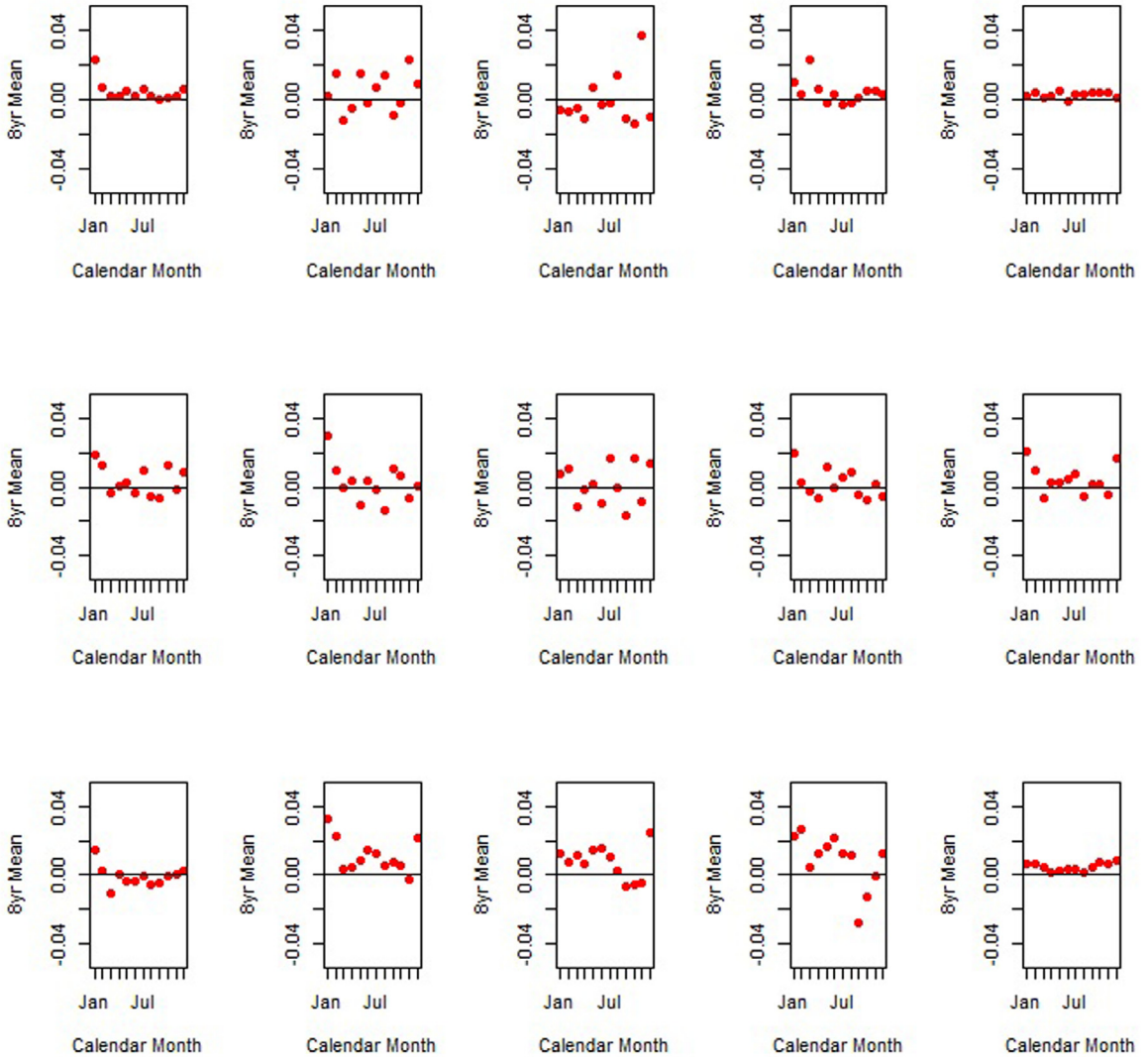

Fig. 4. Plots of twelve monthly means, computed over 8 years (red dots) and 16 years (horizontal lines), for 15 retail series included (from left to right and top to bottom) following the sequence given in the text.

recent work has focused on placing seasonal adjustment diagnostics on a rigorous statistical footing. Following the work of McElroy and Holan (2009), this paper examines the assessment of spectral peaks and incorporates a quantification of Type I error into the decision rule. This advance is achieved by determining the distribution of tapered-spectral estimators and the associated peak functionals, based upon the visual significance criterion of Soukup and Findley (1999). Due to the importance of the seasonal adjustment application, the focus has been on testing at specific seasonal frequencies. However, the methodology is general and can be used for testing at any frequency.

The results presented here are applicable in the context of data that is assumed, under the null hypothesis, to be generated from a stationary process. Although the methods can be applied to nonstationary processes, the results available for studying the behavior of spectral estimates in such situations are limited. Moreover, the lack of a well-defined stationary spectrum for such transformed data makes it harder to define a hypothesis about a spectral quantity such as spectral peak. One could use a model-based (e.g., autoregressive) estimator and derive the properties of the estimators under nonstationarity, but the calculations will be very involved. We have chosen to use the common route of deriving the test based on a stationary assumption, which is then evaluated under a variety of scenarios. We want to emphasize that for the testing framework, only the null model is assumed to be stationary. Empirical power can be analyzed for any non-stationary alternative.

Another important point to note about the results is the apparent conservative nature of the test. As explained in the discussion following Proposition 1 as well as in Sections 4.3 and 4.4, and illustrated by the results of Proposition 1, as well as the numerical values reported in Table 4, the supremum of Type I error probability over the entire boundary of the

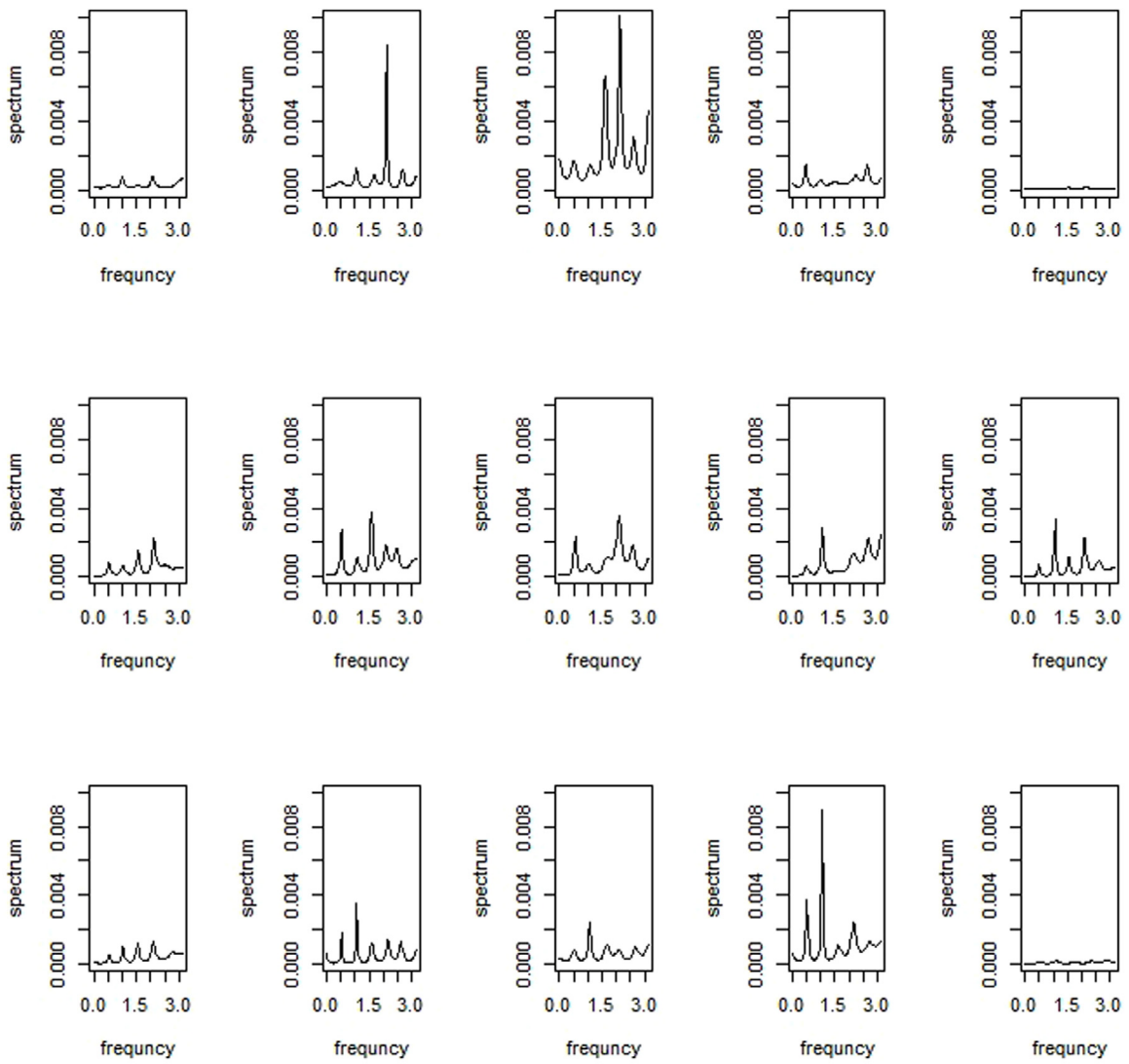

Fig. 5. Plots of spectral density estimates of the 15 data series.

null space achieves the nominal level α; the particular null value where both the left and the right visual significance are small is merely a single point on the boundary of this null space, and the probability there is less than α. In complicated situations such as the nonparametric peak testing problem, it is unreasonable to expect a "similar" test with exact size on the entire boundary of the null region (even in the simple one-sided normal mean problem such tests are not known). If one thinks of absence of peak as the visual significance being equally low on the left and right side, then the null region needs to be redefined as well as the test statistic. In that case one could easily get a size α test for that particular null. However, we have chosen to call a feature a peak only if it is equally pronounced on either side, and any feature that does not conform to that is being declared a "non-peak".

In order to formulate meaningful peak criteria, our research has explored different parametrizations of spectral peaks, and proposed threshold values for peak functionals to properly capture spectral salients at seasonal frequencies. What constitutes a peak and how the human eye perceives a peak are of course complex questions without any definitive answer. We have tried to shed some light on the attributes of the functional peak that should be incorporated in a measure of peak strength. Given our own choices of peak functionals – largely motivated by the prior literature – we have tabulated the empirical size and power for Bartlett-taper spectral estimators, using a fixed bandwidth fraction asymptotic theory for critical values. A chief finding is that eight to ten years of monthly data requires a fairly broad peak functional in order to obtain reasonable power while maintaining the correct size; moreover, for the narrower functional corresponding to the peak functional used in the X-13ARIMA-SEATS software, at least twenty years of monthly data is recommended to obtain the correct proportion of Type I errors. Currently the software uses only eight years of data.

Given that most seasonal adjustments arising from the X-13ARIMA-SEATS software are adequate, and that moreover incorporating statistical uncertainty into the peak measures makes it *harder* to detect a peak, we expect in practice that the frequency of incidences of inadequate adjustment will be reduced with the new method. Essentially, the current visual significance procedure has a very inflated Type I error rate, and generates false detections of residual seasonality. The newer methodology adds greater statistical rigor to the visual significance method, and decreases the incidence of Type I errors.

CRediT authorship contribution statement

Tucker McElroy: Conceptualization, Methodology, Formal analysis, Investigation, Resources, Visualization, Supervision, Project administration, Funding acquisition. **Anindya Roy:** Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Visualization.

Acknowledgments

The authors thank Xiaofeng Shao for stimulating discussions about this problem, and helpful comments from the referees.

Disclaimer

This report is released to inform interested parties of research and to encourage discussion. The views expressed on statistical issues are those of the authors and not necessarily those of the U.S. Census Bureau.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jspi.2020.06.012. Supplementary material available online includes background on seasonal spectral densities and all theoretical proofs. Codes for the paper are available upon request.

References

Bell, W.R., Hillmer, S.C., 1983. Modeling time series with calendar variation. J. Amer. Statist. Assoc. 78, 526–534.
Berger, R.L., 1989. Uniformly more powerful tests for hypotheses concerning linear inequalities and normal means. J. Amer. Statist. Assoc. 84, 192–199.
Findley, D.F., 2005. Some recent developments and directions in seasonal adjustment. J. Off. Stat. 21, 343–365.
Findley, D.F., Lytras, D.P., McElroy, T.S., 2017. Detecting Seasonality in Seasonally Adjusted Monthly Time Series. Research report series, Statistics 2017-03, U.S. Census Bureau.
Grether, D.M., Nerlove, M., 1970. Some properties of optimal seasonal adjustment. Econometrica 38, 682–703.
Hashimzade, N., Vogelsang, T., 2008. Fixed-b asymptotic approximation of the sampling behaviour of nonparametric spectral density estimators. J. Time Series Anal. 29, 142–162.
Hylleberg, S., 1986. Seasonality in Regression. Academic Press, Orlando, Florida.
Liu, H., Berger, R.L., 1995. Uniformly more powerful, one-sided tests for hypotheses about linear inequalities. Ann. Statist. 23, 55–72.
Lytras, D.P., Feldpausch, R.M., Bell, W.R., 2007. Determining aeasonality: a comparison of diagnostics from X-13-ARIMA. In: Proceedings of the Third International Conference on Establishment Surveys. http://www.census.gov/srd/www/sapaper/sapaper.html.
Maravall, A., 2012. Update of seasonality tests and automatic model identification in TRAMO-SEATS. Working paper, Bank of Spain.
McElroy, T., Holan, S., 2009. A nonparametric test for residual seasonality. Surv. Meth. 35, 67–83.
McElroy, T., Politis, D., 2014. Spectral density and spectral distribution inference for long memory time series via fixed-b asymptotics. J. Econometrics 182, 211–225.
McElroy, T., Roy, A., 2017. Detection of Seasonality in the Frequency Domain. Research report series, Statistics 2017-01, U.S. Census Bureau, https://www.census.gov/srd/papers/pdf/RRS2017-03.pdf.
Nerlove, M., 1964. Spectral analysis of seasonal adjustment procedures. Econometrica 32, 241–286.
Parzen, E., 1957. On consistent estimates of the spectrum of a stationary time series. Ann. Math. Stat. 28, 329–348.
Pierce, D.A., 1976. Uncertainty in seasonal adjustment procedures. In: Proceedings of the Business and Economic Statistics Section of the American Statistical Association. Washington, D.C. pp. 528–533.
Pierce, D.A., 1979. Seasonal adjustment when both deterministic and stochastic seasonality are present. In: Zellner, Arnold (Ed.), Seasonal Analysis of Economic Time Series. NBER, pp. 242–280.
Priestley, M.B., 1981. Spectral Analysis and Time Series. Academic Press, New York.
Sasabuchi, S., 1980. A test of a multivariate normal mean with composite hypotheses determined by linear inequalities. Biometrika 67, 429–439.
Soukup, R., Findley, D., 1999. On the spectrum diagnostics used by X-13-ARIMA to indicate the presence of trading day effects after modeling or adjustment. In: Proceedings of the Third International Conference on Establishment Surveys (ICES-III). Montreal, Canada.
Sun, Y., 2014. Let's fix it: Fixed-b asymptotics versus small-b asymptotics in heteroskedasticity and autocorrelation robust inference. J. Econometrics 178, 659–677.
U.S Census Bureau, 2015. X-13ARIMA-SEATS Reference Manual. U.S. Census Bureau, Washington D.C. USA, (Available from https://www.census.gov/ts/x13as/docX13AS.pdf).