ADAPTIVE K-S TESTS FOR WHITE NOISE IN THE FREQUENCY DOMAIN

Hossein Arsham
University of Baltimore
Baltimore, MD 21201, USA

Abstract: Since the classical Kolmogorov-Smirnov (K-S) tests for white noise are completely insensitive to both tails it is difficult in a particular application to detect departure from whiteness. A modified version of the widely used (K-S) test of null hypothesis is constructed, that a given time series is Gaussian white noise, against the alternative hypothesis that the time series contains an added or multiplicative deterministic periodic component of unspecified frequency. The usual K-S test is treated as a special case. The proposed test is more powerful than the ordinary K-S test in detecting extreme (low or high) hidden periodicities. Computational procedures necessary for implementation are given.

AMS Subject Classification: 60H40, 62F25, 60G10
Key Words: Kolmogorov-Smirnov test, frequency analysis, identification of periodic component, goodness-of-fit test for white noise, periodogram, residuals, Spectrum; Stationary process

1. Introduction

A systematic approach in statistical treatments of hydrologic time series often includes testing the goodness of fit of a model. This phase of analysis needs to be checked in order to verify whether it complies with certain assumptions about the model and to verify how well it represents the historical time series. One of the model assumptions to be checked is the independence of residual (Gaussian white noise). If this check is not satisfactory, the model should
be changed and the procedure repeated until a satisfactory model is found. Since the periodogram has simpler sampling properties than auto-correlation coefficients, tests based on the periodogram are considered ideal techniques, particularly for detecting the important frequencies (in problems such as water level oscillations and vibration problems). Therefore, any hypothesis about a mathematical model of the series composition resulting from the periodogram analysis of a time series and the inference of the best fitted mathematical model of dependency of the series is equivalent to testing whether the residuals are from an independent series.

Given a finite realization from a time series $X_t$, we can represent the $n$ observations by the trigonometric polynomial, see e.g. Wilks [6],

$$X_t = \frac{a_0}{2} + \sum_{k=1}^{m} (a_k \cos w_k t + b_k \sin w_k t),$$  \hspace{1cm} (1)

where:

$$w_k = \frac{2\pi k}{n}, \hspace{1cm} k = 0, 1, 2, \ldots, m,$$  \hspace{1cm} (2)

$$a_k = \frac{2}{n} \sum_{t=1}^{n} X_t \cos w_k t, \hspace{1cm} k = 0, 1, 2, \ldots, m,$$  \hspace{1cm} (3)

and

$$b_k = \frac{2}{n} \sum_{t=1}^{n} X_t \sin w_k t, \hspace{1cm} k = 1, 2, \ldots, m,$$  \hspace{1cm} (4)

where $m$ is the smallest integer greater than or equal to $(n - 1)/2$. The periodogram is defined by

$$I_n(w_k) = n \frac{a_k^2 + b_k^2}{2}, \hspace{1cm} k = 1, 2, \ldots, m.$$  \hspace{1cm} (5)

The graph of $I_n(w_k)$ against $k$ presents a highly unstable appearance which makes it difficult for visual inspection. A better method of presenting the information in a periodogram (5) is to plot the cumulative periodogram

$$Y_k = \frac{\sum_{j=1}^{k} I_n(w_j)}{\sum_{j=1}^{m} I_n(w_j)}, \hspace{1cm} k = 1, 2, 3, \ldots, m - 1.$$  \hspace{1cm} (6)
This differs from the periodogram itself, in that the highly unstable behavior of \( I_n(w_k) \)'s is, to large extend, ironed out by the process of accumulation. Thus the cumulative periodogram is a useful diagnostic tool.

It is well known that if \( \{X_t\} \) is Gaussian white noise, then \( Y_k \) random variables will lie close to the \( 45^0 \) line on the graph of \( Y_k \) against \( k \). On the other hand, \( Y_k \) for a process with excess of low frequency will tend to lie above the line. By way of contrast, a process with an excess of high frequency components will tend to have a cumulative periodogram which lies below the line. Therefore, based on cumulative periodogram (6), we shall consider a new procedure which is used to test the null hypothesis \( H_0 \), that the residual \( \{X_1, X_2, \ldots, X_n\} \) is a sequence generated by Gaussian white noise, against the alternative hypothesis that the data is generated by a Gaussian white noise sequence with a superimposed deterministic periodic components.

The basic assumption that the periodic movements in most hydrologic time series constitute deterministic components has both physical justification and practical implication. The astromonic cycles that produce hydrologic periodicities are constant periods in all hydrologic and water resources problems. Amplitudes and phases of individual harmonics usually change when e.g. water passes through a hydrologic environment, by the system's effects on amplification and attenuation of amplitudes, and the shifts of phases. The practical implication of the assumption of deterministic periodic components in e.g. hydrologic series is basically twofold.

the series can be rightfully decomposed into deterministic and stochastic components.

the explanation of components by various physical and/or environmental are causative factors in the regional analysis of series.

Since the periodic component is usually unknown, we will consider tests for hidden periodicities of unspecified frequency. As mentioned earlier, when residuals \( \{X_1, X_2, \ldots, X_n\} \) are independently normally distributed, \( Y_1, Y_2, \ldots, Y_{m-1} \) are distributed like order statistics from a sample of \( m-1 \) independent observation from Uniform \((0, 1)\). The most widely used tests based on periodograms for hidden periodicities is the Kolmogorov-Smirnov (K-S) test, see e.g. Balakrishnan [2], Brockwell and Davis [2] and Stoica and Moses [5]. One problem with the K-S is that it is too wide in the tails therefore it \textit{may not detect} any hidden periodicity at low or high frequencies. The next section provides a new test, which is a modification of the ordinary K-S test, to make it more powerful in detecting extreme hidden periodicities than the ordinary K-S test.
2. A Modified K-S Test

Clearly, the cumulative distribution function with jumps of size \(1/(m-1)\) at \(Y_k, k = 1, 2, ..., m - 1\) is the empirical distribution function of a sample of size \((m - 1)\) from a uniform distribution on \([0,1]\). Therefore, if we plot the \(Y_k\) on a sample distribution function and apply the K-S test of the hypothesis that it is a sample distribution function for a sample of \((m-1)\) obtained from a uniform \([0,1]\) distribution, we have a test of the hypothesis that the original time series is white noise. This testing procedure has been discussed by e.g. Pollock, et. al. [3].

This procedure is precisely equivalent to plotting the standardized, cumulative periodogram

\[
C(X) = \begin{cases} 
0, & X < 1, \\
y, & I \leq X < i + 1, \\
1, & X \geq m,
\end{cases}
\]

for \(i = 1, 2, \ldots, m - 1\),

and rejecting the null hypothesis at the significance level of \(\alpha\) if for any \(X\) in \([1,m]\), the function \(C(X)\) exits from the boundaries

\[
Y = X + K_\alpha (m - 1)^{-\frac{1}{2}},
\]

\[
Y = X - K_\alpha (m - 1)^{-\frac{1}{2}},
\]

where \(K_\alpha\) is the critical value of the ordinary K-S statistics for sample of size \(m-1\) and significance level \(\alpha\). The advantage of this test over others e.g. the Fisher’s test, is that its graphical presentation is universally appealing. However, one problem with the K-S test is that the acceptance bandwidth generated by (8) and (9) is too wide at the both ends. This weakness is caused because the bandwidth is constant for all \(X\).

Therefore, the K-S test may not detect hidden periodicities at either low or high frequencies. In fact the K-S confidence region with \(100(1-\alpha)\%\) confidence must be truncated at both ends, i.e.

\[
P \left\{ \max \left[ X - K_\alpha (m - 1)^{-\frac{1}{2}}, 0 \right] \leq C(X) \leq \min \left[ X + K_\alpha (m - 1)^{-\frac{1}{2}}, 1 \right] \right\} = 1 - \alpha. \tag{10}
\]

That means we may not be able to detect extreme periodicities by using this test.
In the next section we suggest a modification of the K-S acceptance region to remedy this problem. We consider the acceptance regions of the following desirable forms:

For $c > 0$ and $a \geq 0$:

\[ P \{ A_L(x, m) \leq C(x) \leq B_L(x, m), \text{ for all } x \} = 1 - \alpha \]  

(11)

and

\[ P \{ A_R(x, m) \leq C(x) \leq B_R(x, m), \text{ for all } x \} = 1 - \alpha, \]  

(12)

where:

\begin{align*}
A_L(x, m) & = \max \left\{ \frac{X}{(m-1) + ca(m-1)^{\frac{1}{2}}} - a/ [(m-1)^{\frac{3}{2}} + ca], 0 \right\}, \\
B_L(x, m) & = \min \left\{ \frac{X}{(m-1) - ca(m-1)^{\frac{1}{2}}} + a/ [(m-1)^{\frac{3}{2}} - ca], 1 \right\}, \\
A_R(x, m) & = \max \left\{ \frac{X}{(m-1) - ca(m-1)^{\frac{1}{2}}} - a/ [(m-1)^{\frac{3}{2}} - ca], 0 \right\}, \\
B_R(x, m) & 
\end{align*}

(13) to (15)
\[ \text{min} \left\{ \frac{X}{(m-1) + ca(m-1)^{1/2}} + \frac{a}{(m-1)^{1/2} + ca}, 1 \right\}. \] (16)

These particular shapes of acceptance regions were dictated by their analytical tractability and usefulness. Note that region (11) is narrower in the left (this helps to detect hidden low periodicities) and (12) narrower in the right (this helps to detect hidden high periodicities). The factor \( c \) shapes the acceptance region. Larger values of \( c \) imply a region with smaller bandwidth one side (right or left). For \( c = 0 \), the modified regions (11) and (12) reduce to the usual K-S acceptance region (10).

![Figure 2: Desirable acceptance region sensitive to low hidden periodicities with diagonal E](image1)

![Figure 3: Desirable acceptance region sensitive to high hidden periodicities with diagonal E](image2)
In terms of $U_{m-1}$, by an appropriate re-scaling the acceptance region (11) can be expressed as

$$P \left\{ \max \left\{ X \left[ 1 - ac(m-1)^{-\frac{1}{2}} \right] - a(m-1)^{-\frac{1}{2}}, 0 \right\} \leq U_{m-1} \right.$$ 

$$\leq \min \left\{ X \left[ 1 + ac(m-1)^{-\frac{1}{2}} \right] + a(m-1)^{-\frac{1}{2}}, 1 \right\} \right\} = 1 - \alpha. \quad (17)$$

Because $[1 - U_{m-1}(1 - x)]$ and $[1 - U_{m-1}(x)]$ are identical. (12) can be written as:

$$P \left\{ \max \left\{ X \left[ 1 - ac(m-1)^{-\frac{1}{2}} \right] - a(1 + c)(m-1)^{-\frac{1}{2}}, 0 \right\} \leq U_{m-1} \right.$$ 

$$\leq \min \left\{ X \left[ 1 + ac(m-1)^{-\frac{1}{2}} \right] + a(1 + c)(m-1)^{-\frac{1}{2}}, 1 \right\} \right\} = 1 - \alpha. \quad (18)$$

The confidence levels of regions (17) and (18) are related to the crossing probabilities of $U_{m-1}(\cdot)$, a path of the empirical cdf of $m-1$ independent uniform $[0,1]$ random variables.

There are a few algorithms available to compute $\alpha$ using (17). For a complete discussion of these algorithms see Shorack and Wellner [4], Chapter 9. For example in Steck’s algorithm, let

$$G_1(i) = \min \left\{ (i - 1 + a(m-1)^{\frac{1}{2}})/(m-1 - ac(m-1)^{\frac{1}{2}}), 1 \right\},$$

$$i = 0, 1, 2, ..., m - 2, \quad (19)$$

$$G_2(i) = \max \left\{ (i - a(m-1)^{\frac{1}{2}})/(m-1 + ac(m-1)^{\frac{1}{2}}), 0 \right\},$$

$$i = 1, 2, ..., m - 1. \quad (20)$$

Then

$$1 - \alpha = (m-1)\det \frac{[G_1(i) - G_2(j)]_{+}^{j-i+1}}{(j-i+1)!}, \quad (21)$$

where,

$$[G_1(i) - G_2(j)]_{+} = \begin{cases} 
G_1(i) - G_2(j), & \text{if } G_1(i) \geq G_2(j), \\
0, & \text{otherwise,}
\end{cases}$$

and it is understood that all elements of the matrix having subscripts $i > j + 1$ are zero. This algorithm performs well for small size sample. For example for $m = 31$, $c = 1$, and $a = 0.8446$ we get $\alpha = 0.1$, while for $c = 0$, i.e. for
the ordinary K-S version to have the same significance level, we need to set \( a = 1.1916 \).

The asymptotic critical values can be computed by using the theory of weak convergence.

**Theorem.** For any \( a > 0 \) and \( c \geq 0 \) as \( m \to \infty \),

\[
1 - \alpha = 1 + 2 \sum_{k=1}^{\infty} (-1)^k \exp\left(-2k^2a^2(1 + c)\right).
\]  

(23)

**Proof.** Proof follows by the weak convergency, see e.g. Shorack and Wellner [4].

As a numerical example, at \( \alpha = 0.1 \) level for \( c = 0 \), and \( c = 1 \), we obtain \( a = 1.2239, 0.8654 \), respectively.

Usually, the construction of the acceptance region of form (11) and (12) involves calculation of critical value \( a \), given parameters \( m \), \( \alpha \) and a specified value of \( c \).

The choice for numerical values for these parameters naturally depends on the proposed application. One may, however, rely on actual observations in deciding what parameter is suitable. *In practice one constructs several different confidence bounds with different values for parameters \( (a, \text{ and } c) \) using (21), and then selects the one which looks “good” overall. This procedure is subjective.*

### 3. Concluding Remarks

It is often desirable to supplement informal inference from a periodogram with a formal test to compare the periodogram with the theoretical white noise. The ordinary Kolmogorov-Smirnov (K-S) test represents an attempt to uncover any hidden periodicity. One problem with this test is that it may not recognize any hidden periodicity at the extreme frequencies. We modified the ordinary K-S test to be more sensitive to either low or high frequencies. A factor \( c \) shapes the confidence regions. Larger values of \( c \) imply a region with smaller bandwidth in one tail. For \( c=0 \) these regions reduce to the ordinary K-S confidence region.

These bounds are important, as they allow a more realistic way of selecting the desired confidence bounds. Recognizing that analysts are concerned with the risk implication of their decision alternatives, it is suggested that the problem of selecting the parameters of a desirable shape could be based on a
measurable criterion to minimize the risk factor. In certain circumstances a one-sided test may be appropriate. For example, if the alternative hypothesis is that there is an excess of low frequency, only the upper line is relevant. Since an excess of low frequency corresponds to positive serial correlation, such an alternative will often be very reasonable. For the one-sided version of the proposed modified K-S confidence regions, see Arsham [1]. Pseudo codes version of all the procedures presented in this study are available from the author upon request.

Acknowledgements

I am grateful to the reviewers’ useful comments on an earlier version. The National Science Foundation Grant CCR-9505732 supports this work.

References


