# Detecting Periodicity in Nonideal Datasets

R. K. Pearson* H. Lähdesmäki,† H. Huttunen,‡ O. Yli-Harja§

**Abstract**

Many time-series encountered in data mining applications exhibit outliers, missing data, and small sample sizes (specifically, short time-series) that can cause standard methods to perform poorly. Motivated by requirements for the analysis of cDNA microarray time-series data, this paper describes a rank-based approach to periodic component detection for such time-series and examines its performance for several different simulation-based scenarios.

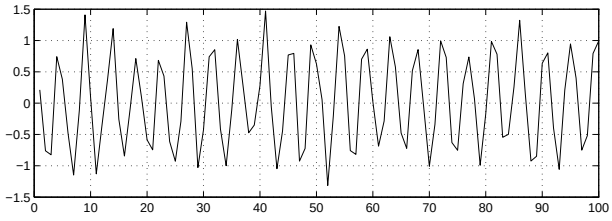


Figure 1: Nominal data sequence: noisy sinusoid

## 1 Introduction

The detection of periodic components in a data sequence is important enough that many different methods have been proposed to solve this problem [6]. Since a periodic component should appear as a peak in a plot of the power spectrum $S_{xx}(f)$ against frequency, one approach is based on the estimation $S_{xx}(f)$ from observed data. The *Blackman-Tukey* (BT) spectrum estimator $\hat{S}_{xx}(f)$ is based on the fact that $S_{xx}(f)$ is the discrete Fourier transform of the autocorrelation function $R_{xx}(m) = E\{x_k x_{k+m}\}$:

$$\hat{S}_{xx}(f) = \sum_{m=-L}^{L} \hat{R}_{xx}(m) e^{-i2\pi m f T}. \tag{1.1}$$

The autocorrelation function is typically estimated as

$$\hat{R}_{xx}(m) = \frac{1}{N-m} \sum_{k=1}^{N-m} x_k x_{k+m}. \tag{1.2}$$

In practice, the required values of $\hat{R}_{xx}(m)$ for $m < 0$ are obtained by symmetry: $R_{xx}(-m) = R_{xx}(m)$, so that $\hat{R}_{xx}(m)$ need only be computed for $m = 0, 1, 2, \ldots, L$. Also, because the amount of data on which $\hat{R}_{xx}(m)$ is based decreases with increasing $m$, it is standard practice to limit the number of autocorrelation lags $L$ used in this estimator. Here, we take $L = N/4$, a rule of thumb that has been suggested for time-series of at least moderate length (e.g., $N > 50$). To keep our results as comparable as possible, we adopt this rule here, even for the short time-series considered in Sec. 6.

This paper considers three variations of the estimator defined in Eqs. (1.1) and (1.2), all based on replacement of $\hat{R}_{xx}(m)$ with an alternative, aimed at addressing some of the forms of nonideal sequence behavior described in the next section.

## 2 Nonideal data sequences

Because data mining generally deals with data collected under relatively uncontrolled and uncontrollable conditions, these datasets often exhibit various forms of nonideal behavior. This section briefly discusses three types of nonideal behavior: outliers, missing data, and short time-series. Fig. 1 shows a 100 point sequence of samples of a noisy sinusoid, which may be regarded as the "ideal" sequence we would like to have available for analysis in the simulation examples considered here. More specifically, the sinusoid in this sequence is of frequency $f \simeq 0.22$ and unit amplitude, contaminated additively with a zero-mean, Gaussian white noise sequence with standard deviation $\sigma = 0.2$.

### 2.1 Outliers

*Outliers* are data points that are clearly inconsistent with most of the other points in the dataset. A typical example is shown in Fig. 2, where the noisy sinusoidal time-series shown in Fig. 1 has been contaminated by randomly replacing approximately 15% of the dataset with a single point that lies 8 standard deviations (of the sinusoidal signal alone) from the mean of the original dataset. The result is the presence of the pronounced spikes clearly evident in Fig. 2.

*Daniel Baugh Institute for Functional Genomics and Computational Biology, Thomas Jefferson University, Philadelphia, USA

†Tampere International Center for Signal Processing, Tampere University of Technology, Tampere, Finland.

‡Tampere International Center for Signal Processing, Tampere University of Technology, Tampere, Finland.

§Tampere International Center for Signal Processing, Tampere University of Technology, Tampere, Finland.

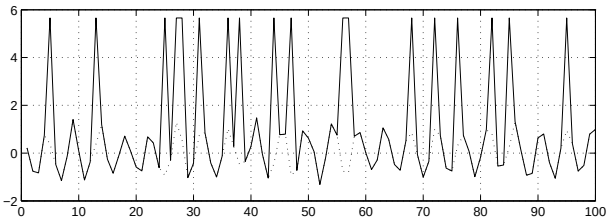

Figure 2: Noisy sinusoid with 15% random outliers

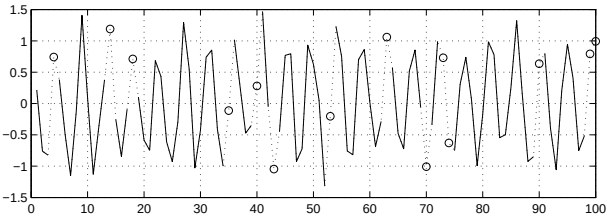

Figure 3: Noisy sinusoid with randomly missing samples

**2.2 Missing data** It has been suggested that in many applications, as much as $\sim$ 30% missing data is typical [3]. Fig. 3 shows the noisy sinusoidal data sequence from Fig. 1 with approximately 15% of the samples missing, indicated by the open circles.

The Blackman-Tukey estimator can be modified to account for missing data at known sampling times $k$ by replacing $\hat{R}_{xx}(m)$ with the following modified estimator. First, for each integer $0 \leq m \leq L$, define $I_m$ to be the set of all time indices $k$ for which both $x_k$ and $x_{k+m}$ are available, and define $K_m$ as the number of elements in the set $I_m$. So long as $K_m \neq 0$, one possible unbiased estimate of $R_{xx}(m)$ is given by

$$\tilde{R}_{xx}(m) = \frac{1}{K_m} \sum_{k \in I_m} x_k x_{k+m}. \tag{2.3}$$

In the absence of missing data, this expression reduces to the standard estimator given in Eq. (1.2). Conversely, if $K_m = 0$, the pattern of missing data is such that $R_{xx}(m)$ cannot be estimated at all, a situation that can occur when data values are *systematically* missing.

**2.3 Short time series** Data mining is fundamentally concerned with the analysis of large datasets, but one special case that is becoming increasingly important is the problem of analyzing large collections of short time-series. In particular, high-throughput methods in functional genomics studies frequently lead to large collections (e.g., thousands or tens of thousands) of short time-series, typically of length $N \sim 10$ to $\sim 20$. In many of these experiments, the anticipated state-of-the-art for the forseeable future simply does not permit the generation of meaningful time-series of length $N \gtrsim 100$, which would be preferred for detection of periodicities and other types of time-series analysis. Practical consequences of dealing with very short time-series are first, the inapplicability of several of the more sophisticated methods and second, a general increase in the uncertainty of computed results since variability generally decreases with increasing sample size.

## 3 Rank-based extensions

This paper considers the extension of a recently proposed, rank-based alternative autocorrelation estimator [4] to problems of spectrum estimation. This estimator is a moving-window extension of the Spearman rank correlation coefficient [1, p. 61], quantifying the association between the sequences $\{x_k\}$ and $\{x_{k+m}\}$. The resulting quantity, $\rho^S_{xx}(m)$ is actually an alternative estimator of the correlation coefficient $\rho_{xx}(m)$ between these sequences, which is related to the autocorrelation function $R_{xx}(m)$ by $R_{xx}(m) = \bar{x}^2 + \sigma_x^2 \rho_{xx}(m)$, where $\bar{x}$ is the mean of the sequence and $\sigma_x^2$ is its variance. Since it is important to remove the mean of the sequence prior to spectrum estimation to avoid low frequency artifacts and since $\sigma_x^2$ is simply a scale factor, the problem of detecting periodic components in a data sequence may equally well be based on $\rho_{xx}(m)$ as $R_{xx}(m)$. Consequently, we consider spectral estimators of the form

$$\tilde{S}_{xx}(f) = \sum_{m=-L}^{L} \tilde{\rho}_{xx}(m) e^{-i2\pi m f T}, \tag{3.4}$$

where $\tilde{\rho}_{xx}(m)$ estimates the correlation coefficient between $\{x_k\}$ and $\{x_{k+m}\}$. More specifically, with $I_m$ and $K_m$ defined as in Sec. 2.2, we consider the missing data-adapted correlation coefficient between the *data ranks* $R_x(i)$ and $R'_x(i)$, defined by:

$$\rho^S_{xx}(m) = \frac{12}{K_m(K_m^2 - 1)} \sum_{i \in I_m} \left( R_x(i) - \frac{K_m + 1}{2} \right) \left( R'_x(i) - \frac{K_m + 1}{2} \right). \tag{3.5}$$

Here, $R_x(i)$ denotes the rank of $x_i$ in the set $S = \{x_j | j \in I_m\}$ and $R'_x(i)$ denotes the rank of $x_{i+m}$ in the set $S' = \{x_{j+m} | j \in I_m\}$. Because both of these rank sequences assume every value from 1 to $K_m$ precisely once, their average is $(K_m + 1)/2$, independent of the data values $\{x_k\}$, and their variance can be shown to be $K_m(K_m^2 - 1)/12$ [1, p. 89]. More generally, since $\rho^S_{xx}(m)$ is the correlation coefficients between ranks, it is bounded by $-1 \leq \rho^S_{xx}(m) \leq 1$ for all $m$.

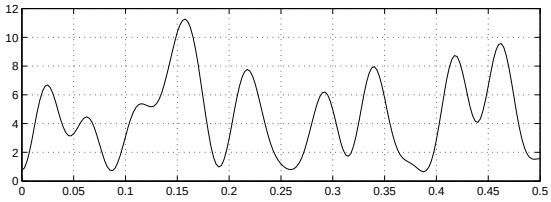


Figure 4: BT spectrum estimate, 15% contamination

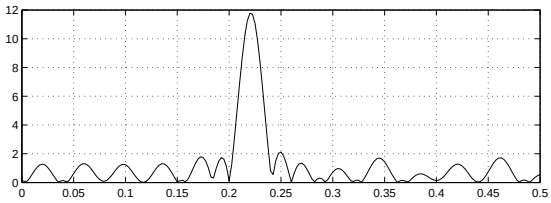


Figure 5: Rank-based spectrum, 15% contamination

## 4 Results for contaminated series

Results using the classical Blackman-Tukey method are shown in Fig. 4 for a sequence containing 15% outliers, all lying 8 standard deviations above the uncontaminated signal mean. These results clearly illustrate the severe outlier sensitivity of the BT method: although there is a peak in the estimated spectrum at approximately the correct value ($f = 0.22$), there are also four other peaks of greater magnitude at frequencies $f \simeq 0.16$, $f \simeq 0.34$, $f \simeq 0.42$, and $f \simeq 0.46$, along with peaks that are only slightly less intense at $f \simeq 0.02$ and $f \simeq 0.29$. The rank-based spectrum estimate introduced in Sec. 3 is shown in Fig. 5 for the same data sequence and the results are exactly what we would hope to see: a single, well-defined peak is clearly evident, centered at $f \simeq 0.22$. As in the BT example, the estimated spectrum also contains a number of other peaks, but here these secondary peaks are smaller than the main peak by a factor of 6 or more.

Although these results indicate a dramatic difference between methods, it is important to ask how representative these differences are. To address this question, Figs. 6 and 7 present spectral envelope plots, which may be viewed as dynamic extensions of boxplots [2, p. 44], consisting of six overlaid curves of spectral intensity vs. frequency. These plots summarize the results of 100 independent simulations, based on sequences with the same amplitude sinusoid, the same Gaussian white noise level, and contaminated by a specified fraction of randomly spaced outliers, all of the same intensity. The six curves in these plots represent:

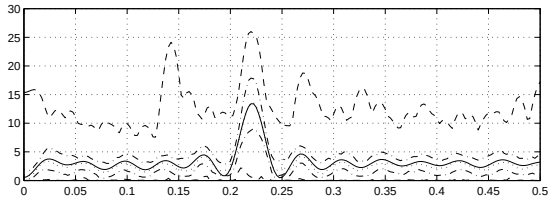


Figure 6: Range of 100 BT spectra, 10% outliers

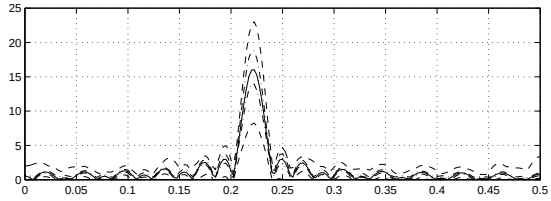


Figure 7: Range of 100 rank-based spectra, 10% outliers

1. the minimum of the 100 values at frequency $f$,
2. the lower quartile (the smallest value not exceeded by 25% of the values at frequency $f$),
3. the median of the 100 values at frequency $f$,
4. the mean of the 100 values at frequency $f$,
5. the upper quartile (the smallest value not exceeded by 75% of the values at frequency $f$),
6. the maximum of the 100 values at frequency $f$.

In these plots, the mean spectrum (curve number 4) is shown as a solid line, the minimum and maximum values are shown as dashed lines, the upper and lower quartiles are shown as dash-dotted lines, and the median is shown as a dotted line.

Fig. 6 summarizes the results obtained with the classical BT estimator for 100 time-series, each of length $N = 100$ and each contaminated with 10% outliers, randomly distributed throughout the data sequence. The solid curve in the center of the plot represents the mean of the 100 simulations, which is seen to be generally consistent with our expectations. This consistency is somewhat misleading, however, because it is the average of 100 individual spectrum estimates, thus representing an estimate based on the $10,000$ total data points in the 100 simulations. What is clear from the other curves in this plot is that the individual sample-to-sample variability of the spectra can be enormous, with extreme members of this family exhibiting many spurious details, exactly as seen in Fig. 4.

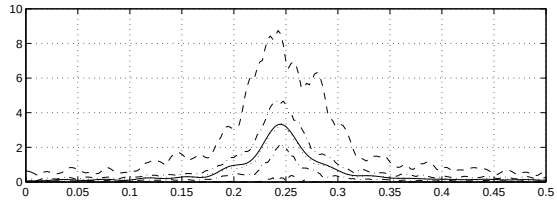


Figure 8: Range of 100 BT spectra, 15% *unacknowledged* missing data

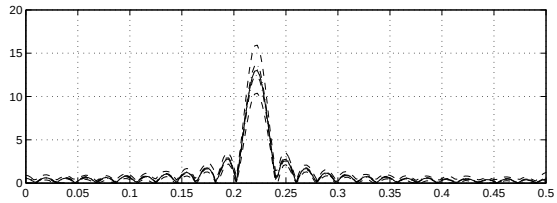


Figure 9: Range of 100 BT spectra, *modified for* 15% *missing data*

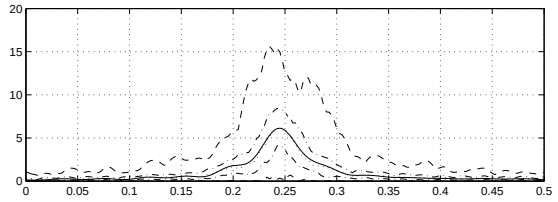


Figure 10: Range of 100 rank-based spectra, 15% *unacknowledged* missing data

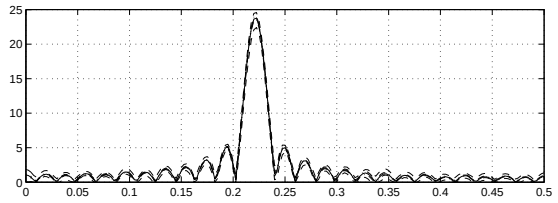


Figure 11: Range of 100 rank-based spectra, *modified for* 15% *missing data*

For comparison, Fig. 7 shows the results for the same 100 simulated data sequences, obtained using the rank-based spectrum estimation procedure. Although its height varies somewhat, the observed peak at $f \simeq 0.22$ is always the dominant spectral feature, in dramatic contrast to the results shown in Fig. 6 for the classical BT estimate. Taken together, Figs. 6 and 7 illustrate that the individual sample results presented in Figs. 4 and 5 are indeed representative.

## 5 Results with missing data

Fig. 8 shows the range of 100 estimated spectra using the classical Blackman-Tukey estimator, making no allowance for the irregular spacing induced by the 15% missing data in the samples from which these spectra were estimated. In contrast to the outlier-contaminated BT results discussed in the previous section, the spectral envelopes shown in Fig. 8 may generally be characterized as describing a single spectral peak, centered at roughly the correct frequency, but with two significant differences. First and most obviously, the peak seen in the average BT spectrum is much broader than that seen in the prevous examples, and somewhat irregular in shape. The second point is that the center of the peak in the average spectrum is shifted to a higher frequency than that of the periodic component in the data sequence ($f \simeq 0.25$ vs. $f \simeq 0.22$). In contrast, Fig. 9 shows the results for the same 100 data sequences, obtained using the modified version of the Blackman-Tukey estimator that correctly accounts for the data points that are missing from the sequence. Here, the results are much better, with all of the estimated spectra exhibiting a single, well-defined peak centered at approximately the frequency of the periodic component that is present in the data sequence.

Fig. 10 shows the results analogous to those plotted in Fig. 8, but for the unmodified rank-based spectral estimator. Very close examination of these two figures reveals that there are some differences, but generally quite minor ones. Similarly, Fig. 11 shows the results obtained using the rank-based spectrum estimator introduced in Sec. 3, modified for missing data as discussed for the BT estimator in Sec. 2.2. As in the case of the modified BT estimator, correctly accounting for missing data gives a *much* better spectral estimate, exhibiting the expected single, narrow peak, centered at $f \simeq 0.22$. In fact, careful comparison of Figs. 11 and 9 reveals that the modified rank-based estimator exhibits a peak that is both somewhat more pronounced relative to its side-lobes than the modified BT estimator, and much less variable in terms of its peak heights.

## 6 Results for short time series

Any consistent estimator becomes increasingly well-behaved with increasing sample size. It follows as a corollary, then, that we can expect poorer performance for the estimators considered here when we apply them to short time-series than the results just presented for time-series of moderate length. The following

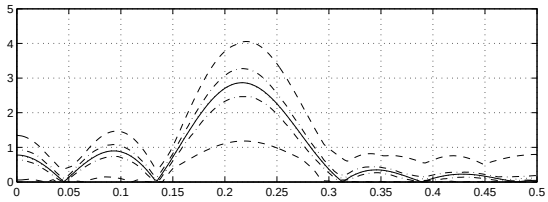


Figure 12: Range of 100 modified BT spectra, 15% missing data

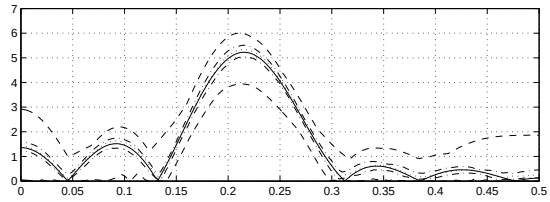


Figure 13: Range of 100 modified rank-based spectra, 15% missing data

discussion briefly illustrates the nature and magnitude of this breakdown by considering the performance of the spectral estimators discussed here for time-series of length $N = 20$, corresponding roughly to the length of published yeast microarray time-series [5]. In addition, because these and other microarray time-series usually exhibit irregular spacing, we focus here on the case of short time-series with missing data.

Fig. 12 summarizes the results obtained using the modified Blackman-Tukey estimator applied to 100 time-series of length $N = 20$ with 15% randomly missing data. Relative to Fig. 9, which summarized the corresponding results for time-series of length $N = 100$, this figure exhibits three primary differences, all of which are due to the smaller sample size. First, the width of the observed spectral peak is much wider, although it remains centered at the correct frequency, $f \simeq 0.22$. The second difference is the much greater variability in the height of this peak among the individual spectral estimates. Finally, the third significant difference is the much lower height of this main spectral peak, relative to the adjacent side-lobes.

Fig. 13 shows the results obtained using the rank-based estimator, modified to correctly account for the 15% missing data in the sample. As with the longer time-series, the results obtained with this estimator are more consistent than those obtained using the BT estimator, but the same three general conclusions noted for the BT estimator apply here as well: the main peak is broader, more variable, and lower relative to the side-lobes than in the results obtained for the longer time-series. Again, this is simply a consequence of attempting to analyze shorter time-series: the problem becomes harder and the results become poorer.

## 7 Summary

This paper has introduced a simple alternative to the classical Blackman-Tukey (BT) spectral estimator based on rank correlations, motivated by the need for a simple spectrum estimation procedure that is relatively resistant to such non-ideal data features as outliers and missing data [4]. The results presented here demonstrate the clear superiority of the rank-based approach over the classical BT estimator in the presence of outliers. Simple extensions of both the classical and rank-based estimators are introduced to deal with missing data, and it is shown first, that such extensions are important and second, that the rank-based estimator again generally yields better spectral estimates than the classical estimator. Finally, the extent to which all of the methods considered here degrade for short time-series is investigated and it is seen that, for time-series as short as $N = 20$, the rank-based approach generally retains its advantage over the classical BT estimator in detecting periodic components in a data sequence.

Because space limitations here did not permit detailed discussions of both the rank-based method and its application to a real microarray dataset, we have chosen to describe the method and summarize its performance for simulation-based examples for which the correct results are known. A paper describing the application of this method to microarray time-series data is currently in preparation.

## References

[1] J. Hájek, Z. Sidák, and P.K. Sen, *Theory of Rank Tests*, Academic Press, New York, 1999.

[2] D.C. Hoaglin, F. Mosteller, and J.W. Tukey, *Fundamentals of Exploratory Analysis of Variance*, Wiley, New York, 1991.

[3] M. Lloyd-Williams, *Empirical studies of the knowledge discovery approach to health-information analysis*, in Knowledge Discovery and Data Mining, Institution of Electrical Engineers, London, 1999.

[4] R.K. Pearson, *Rank-based auto- and cross-correlations*, submitted for publication.

[5] P.T. Spellman *et al.*, *Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast* Saccharomyces cerevisiae *by Microarray Hybridization*, Molecular Biol. Cell, 9 (1998), pp. 3273–3297.

[6] P. Stoica, *List of references on spectral line analysis*, Signal Proc., 31 (1993), pp. 329–340.