A6523
Signal Modeling, Statistical Inference and Data Mining in Astrophysics
Spring 2011

Reading
Chapter 8 “Maximum Entropy Probabilities”

Lecture 12
Aspects of FT based spectral analysis
Variations on the theme

Projects: need to choose a topic
• can be theoretical, numerical simulation, data processing
• e.g. explore an algorithm and underlying statistics
Power Spectrum of a Random Process Passed Through a Linear Filter

\[ X(t) \rightarrow h(t) \rightarrow Y(t) \]

\[ Y(t) = h(t) * X(t) = \int dt' X(t - t') h(t') \]

Now find the ACF of \( Y(t) \):

\[ \langle Y^*(t)Y(t + \tau) \rangle \equiv \langle \int dt' \int dt'' X^*(t - t')X(t + \tau - t'')h^*(t')h(t'') \rangle \]

\[ = \int\int dt' dt'' \langle X^*(t - t')X(t + \tau - t'') \rangle h^*(t')h(t'') \]

We have

\[ R_Y(\tau) = h^*(-\tau) * h(\tau) * R_X(\tau) \]

and the spectrum is

\[ S_y(\omega) = \tilde{H}^*(\omega)\tilde{H}(\omega)S_x(\omega) \]

Therefore

\[ S_Y(\omega) = |H(\omega)|^2 S_X(\omega) \]
Notes:

1. For linear filters with impulse responses $h(t)$ that have unit area,

$$\int dt \ h(t) \equiv \tilde{H}(0) = 1$$

then

$$\langle Y \rangle = \langle X \rangle$$

and the variance of $Y$ will be less than or equal to the variance of $X$:

$$\sigma_y^2 \leq \sigma_x^2$$

equality when $h(t) = \delta(t)$

because

$$\sigma_x^2 - \sigma_y^2 = \langle x^2 \rangle - \langle y^2 \rangle = \frac{1}{2\pi} \int d\omega \ S_x(\omega)[1 - |H(\omega)|^2] \geq 0$$

2. e.g. Low pass filters: $|H(\omega)| \leq \tilde{H}(0) = 1$

Low pass filters that preserve the mean reduce the variance; they smooth the input.

In general,

$$\frac{\sigma_Y}{\langle Y \rangle} \leq \frac{\sigma_X}{\langle X \rangle}$$

for low pass filters
A Third Moment

Bispectrum: Let $\gamma_X(t_1, t_2, t_3) = \langle X(t_1)X(t_2)X(t_3) \rangle$. If stationary to third order, then $\gamma_X$ depends only on

\begin{align*}
    t_2 - t_2 & \equiv \tau_1 \\
    t_3 - t_1 & \equiv \tau_2 \\
    t_3 - t_2 & \equiv \tau_2 - \tau_1
\end{align*}

For a WSS process

$$
\gamma_X = \gamma_X(\tau_1, \tau_2)
$$

Now Fourier transform (2D) to find the bispectrum:

$$
S(\omega_1, \omega_2) = \int d\tau_1 \ e^{-i\omega_1 \tau_1} \int d\tau_2 \ e^{-i\omega_2 \tau_2} \cdot \gamma_X(\tau_1, \tau_2)
$$

It is the distribution of the third moment in frequency space because

$$
\langle x^3(\tau) \rangle = \frac{1}{(2\pi)^2} \iint d\omega_1 \ d\omega_2 \ S(\omega_1, \omega_2)
$$

The bispectrum:

1. contains phase information which the power spectrum does not
2. is useful in studies of time asymmetries of a random process
3. used in studying ocean waves
4. image formation in speckle interferometry (recall similarity of power spectra (second moments) of different kinds of music).
Periodogram

• The spectral estimator we have investigated is

\[ \hat{S}_T(f) = \frac{1}{2T} |\tilde{X}_T(f)|^2 \]

\[ X_T(f) = \int_{-T}^{T} dt \, x(t) e^{-2\pi if t} \]

• This is called the periodogram and represents one approach to spectral estimation
Summary of FT-based Spectral Estimators

- The spectral estimator based on $|FT|^2$ does not converge to the ensemble mean spectrum
  - based on a single FT of length $2T$
- The problem is that the estimator yields only two degrees of freedom per frequency.
  - true for continuous $t,f$ case
  - true for discrete case (DFT)
- Estimation errors have a chi-square pdf with 2 dof that is a one-sided exponential function
  - The good news: we know the PDF and can calculate false-alarm probabilities, etc.
  - The bad: tail of PDF much longer than for a Gaussian with the same rms value
  - higher false alarm rate for a given threshold $p_{\text{min}} = <p> + m\sigma$
Fixes to the Estimation Error Problem

• Increase the number of DoF:
  – smooth the spectrum by convolving with a low-pass filter
    • if smoothing is by $N_s$ frequency samples, error decreases as $1/\sqrt{N_s}$
    • ok if structure in spectrum is broader than $N_s$
    • features narrower than $N_s$ will have reduced amplitudes
      – $\rightarrow$ loss of sensitivity
      – $\rightarrow$ tradeoff between statistical errors and resolution
  
• there is no “correct” approach: it depends on what you are seeking from the data, the nature of the signal, etc.
Fixes to the Estimation Error Problem (2)

• Increase the number of DoF:
  – sum L independent spectral estimates
    • estimation errors $\sim 1 / \sqrt{L}$
    • requires that L x 2T data are available
  – or take the original data span [0, 2T] and divide it into shorter blocks:
    • [0, 2T/L], [2T/L, 2x2T/L], … [(L-1)2T/L, 2T]
    • then calculate shorter FTs.
      – $\rightarrow$ less resolution (again)
    • essentially the same as spectral smoothing
Window Carpentry

• Spectral smoothing:
  – smoothing functions = windows

• Convolution in the frequency domain = multiplication in the time-lag domain:
  – recall
    \[ \langle x(t)x^*(t + \tau) \rangle \leftrightarrow S(f) = \langle |\tilde{X}(f)|^2 \rangle \]
  – So we define lag and spectral windows
    \[ w(\tau) \leftrightarrow \tilde{W}(f) \]
    and then
    \[ \langle x(t)x^*(t + \tau) \rangle \times w(\tau) \leftrightarrow S(f) = \langle |\tilde{X}(f)|^2 \rangle \ast \tilde{W}(f) \]

These are ensemble average expressions, but the same applies to spectral estimators
Window Carpentry:

- We can extend the diagram we had before:

  \[ X(t) \xrightarrow{\text{ACF}} \hat{R}_x(\tau) \xrightarrow{\text{FFT}} \hat{S}_x(f) \]

  \[ W(\tau) = \text{128 window} \xrightarrow{\text{FFT}} \hat{R}_{w_x}(\tau) \xrightarrow{\text{spectral window}} \hat{S}_{w_x}(f) \]

- Path 1 is more efficient than 2 if all lags of the ACF are calculated ($\sim N \log_2 N$ vs $N^2$)
- But there are limitations on the bandwidth over which an FFT can be computed in real time (getting better all the time)
- Path 2 can be made very efficient for gaussian processes by quantizing the signal to $\leq 2$ bits
  - $\rightarrow$ correlation spectrometers
### TABLE 13-1

<table>
<thead>
<tr>
<th></th>
<th>( w(t) )</th>
<th>( W(\omega) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Bartlett</td>
<td>( 1 -</td>
<td>t</td>
</tr>
<tr>
<td>2. Tukey</td>
<td>( \frac{1}{2}(1 + \cos \pi t) )</td>
<td>( \frac{\pi^2 \sin \omega}{\omega(\pi^2 - \omega^2)} )</td>
</tr>
<tr>
<td>3. Parzen</td>
<td>( \frac{\pi^2}{2} )</td>
<td>( \frac{1}{2} \left( \frac{\sin \omega/4}{\omega/4} \right)^4 )</td>
</tr>
<tr>
<td>4. Papoulis†</td>
<td>( \frac{1}{\pi}</td>
<td>\sin \pi t</td>
</tr>
</tbody>
</table>


---

13-2 SPECTRAL ESTIMATION 451
Other Issues

• So far, precision vs. resolution
• Other issues include
  – cases with nonuniform sampling
  – spectral leakage (sidelobes)
  – bias
  – a priori information: how to include?
  – goal of the spectral analysis?
• The entire set of issues has motivated development of alternative methods, including:
  – Lomb-Scargle (Fourier) method to handle non-uniform sampling
  – CLEAN algorithm (removes window function from spectrum, including sidelobes)
  – Multiple taper methods
  – Maximum entropy estimation
  – Bayesian methods
• While the resulting spectral estimates look quite different (mathematically) and have different properties, the periodogram is often lurking in the background
A Simple Example

• Consider white noise that we calculate a spectral estimate for using the DFT.
  – N data points, \( \{x_j, j=1,N\} \)
  – what does the spectral estimate look like?

• Why would we use such an estimator?

• Suppose we know that the process is white noise. What would be a better spectral estimator than the DFT based one?

• In general, where the “true” spectrum could be anything, how do we incorporate prior information?
Lomb-Scargle Method

• Two main features of the periodogram for uniformly distributed data:
  – the exponential PDF for spectral amplitudes
  – the independence of the spectral amplitudes $S(f)$ and $S(f')$ as $T \to \infty$

• The second property is related to the fact that $e^{i\omega t}$ for the $\omega$ of interest are a set of orthogonal basis functions

• With nonuniform sampling, orthogonality no longer applies
Scargle’s Approach

The periodogram is then conventionally defined as

\[ P_X(\omega) = \frac{1}{N_0} |\text{FT}_X(\omega)|^2 \]

\[ = \frac{1}{N_0} \left| \sum_{j=1}^{N_0} X(t_j) \exp(-i\omega t_j) \right|^2 \]

\[ = \frac{1}{N_0} \left[ \left( \sum_j X_j \cos \omega t_j \right)^2 + \left( \sum_j X_j \sin \omega t_j \right)^2 \right] \]

A redefined form preserves the properties of the uniformly sampled case:

\[ P_X(\omega) = \frac{1}{2} \left\{ \frac{\left( \sum_j X_j \cos \omega (t_j - \tau) \right)^2}{\sum_j \cos^2 \omega (t_j - \tau)} + \frac{\left( \sum_j X_j \sin \omega (t_j - \tau) \right)^2}{\sum_j \sin^2 \omega (t_j - \tau)} \right\}, \]

\[ \tau = (1/2\omega) \tan^{-1} \left[ \frac{\sum_j \sin 2\omega t_j}{\sum_j \cos 2\omega t_j} \right]. \]
LS Spectral Analysis

• Lomb got the same result by considering least-squares fitting of sinusoids

• The same issues of sidelobes, statistical errors still apply and have to be dealt with

• The attractiveness of the LS method is that it yields well-understood features of the spectral estimator

• There is an advantage to nonuniform sampling: aliasing is related to the shortest time interval between samples
Window functions derived from the time-series sampling for astrometric data of stars equal vs unequal sampling

Fig. 3.—Logarithmic plots of the classical periodogram windows obtained from astrometric data for the stars: (a) G96–45, (b) G146–72, and (c) Wolf 1062. In each case, the solid line is the window calculated from the classical formulas (D3) and (D4), while the dotted line shows for comparison the window function for even spacing with the same total time interval and number of data points. The open octagonal symbols are the window function evaluated at the grid of frequencies defined in eq. (D1), while the solid curve is oversampled by a factor of 10 relative to this grid.
Gene expression

Detecting periodic patterns in unevenly spaced gene expression time series using Lomb–Scargle periodograms

Earl F. Glynn¹, Jie Chen²,³,* and Arcady R. Mushegian⁴,³

¹Stowers Institute for Medical Research, 1000 East 50th Street, Kansas City, MO 64110, USA, ²Department of Mathematics and Statistics, University of Missouri-Kansas City, 5100 Rockhill Road, Kansas City, MO 64110, USA and ³Department of Microbiology, Immunology and Molecular Genetics, University of Kansas Medical Center, Kansas City, KS 66160, USA

Received on June 13, 2005; revised on October 26, 2005; accepted on November 15, 2005
Advance Access publication November 22, 2005
Associate Editor: Joaquín Dopazo

ABSTRACT

Motivation: Periodic patterns in time series resulting from biological experiments are of great interest. The commonly used Fast Fourier Transform (FFT) algorithm is applicable only when data are evenly spaced and when no values are missing, which is not always the case in high-throughput measurements. The choice of statistic to evaluate the significance of the periodic patterns for unevenly spaced gene expression time series has not been well substantiated.

Methods: The Lomb–Scargle periodogram approach is used to search time series of gene expression to quantify the periodic behavior of every gene represented on the DNA array. The Lomb–Scargle periodogram analysis provides a direct method to treat missing values and unevenly spaced time points. We propose the combination of a Lomb–Scargle test statistic for periodicity and a multiple hypothesis testing procedure with controlled false discovery rate to detect significant periodic gene expression patterns.

Results: We analyzed the Plasmodium falciparum gene expression dataset. In the Quality Control Dataset of 5080 expression patterns, we found 4112 periodic probes. In addition, we identified 243 probes with periodic expression in the Complete Dataset, which could not be examined in the original study by the FFT analysis due to an excessive number of missing values. While most periodic genes had a period of 48 h, some had a period close to 24 h. Our approach should be applicable for detection and quantification of periodic patterns in any unevenly spaced gene expression time-series data.

Availability: The computations were performed in R. The R code is available from http://research.stowers-institute.org/efg/2005/LombScargle

Contact: cheni@umkc.edu

Supplementary information: The online supplement is available at http://research.stowers-institute.org/efg/2005/LombScargle

cell division (Mitchison, 2003), circadian rhythms (Crosthwaine, 2004; Prolo et al., 2005), morphogenesis of periodic structures, such as somites in vertebrates (Dale et al., 2003), complex life cycles of some microorganisms (Lakin-Thomas, 2004; Roversi et al., 2005) and many others.

With the help of gene expression technology, biologists can study the mechanisms that control a particular biological rhythm more closely. For instance, to study how the major oscillator in the suprachiasmatic nuclei (SCN) and in the liver regulates behavioral and physiological rhythms in the whole organism, Panda et al. (2002) used high-density oligonucleotide arrays to measure gene expression in the mouse tissue samples taken every 4 h during two complete circadian cycles and applied a cosine wave-fitting algorithm (Harmer et al., 2000) to identify clusters of circadian-regulated genes among more than 7000 genes. They found that about 650 cycling transcripts were under circadian regulation specific to either the SCN or the liver.

As with many other types of high-dimensional data, the choice of algorithm and statistic to identify significantly periodic patterns of gene expression is a challenge (Wichert et al., 2004). Zhao et al. (2001) used a single-pulse model (SPM) to identify periodic transcripts in the Saccharomyces cerevisae yeast microarray data based on the assumption that the cell-cycle-regulated transcripts will peak only once per cycle. Bar-Joseph et al. (2003) developed an algorithm to represent time series of gene expression as continuous curves using a cubic spline method and used this algorithm to estimate missing values in time series. Langmead et al. (2003) proposed an algorithm that uses autocorrelation to perform linear-time search in frequency and phase, and then use the undirected Hausdorff distance as a similarity measure to cluster genes of similar cyclic patterns together. Johansson et al. (2003) used a multidimensional scaling heat source evaporation model to identify cell