What is Spectral Analysis?

- one of most widely used (& lucrative!) methods in data analysis
- can be regarded as
  - analysis of variance of time series using cosines & sines
  - cosines & sines + statistics (or Fourier theory + statistics)
- today’s lecture: introduction to spectral analysis
  - notion of a ‘time’ series
  - $0.25$ introduction to time series analysis, with some basic notions from ‘time domain’ analysis (subject of Stat 519)
  - definition of simplified version of spectrum and two methods for estimating (nonparametric and parametric)
  - see Chapter 1 for details

Time Series

- what is a time series?
  - ‘one damned thing after another’ (R. A. Fisher?)
  - denote by $x_t, t = 1, \ldots, N$
  - four examples, each with $N = 128$ (Figs. 2 & 3 in textbook)
First Example: Wind Speed Time Series

Second Example: Atomic Clock Time Series
Third Example: Willamette River Time Series

Fourth Example: Ocean Noise Time Series
Time Series Analysis

- goal of time series analysis:
  - quantify characteristics of time series
- sample mean & variance (two well-know statistics)
  \[ \bar{x} = \frac{1}{N} \sum_{t=1}^{N} x_t \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{N} \sum_{t=1}^{N} (x_t - \bar{x})^2, \]

  capture univariate properties, but do not capture bivariate properties, i.e., do not tell us how \( x_t \) and \( x_{t+k} \) are related

Lagged Scatter Plots: I

- tell us about bivariate distribution of separated pairs
- \( x_{t+1} \) versus \( x_t, \ t = 1, \ldots, N - 1 \): lag 1 scatter plot
- four examples (Fig. 4)
Lag 1 Scatter Plot for Willamette River Series

Lag 1 Scatter Plot for Ocean Noise Series
Lagged Scatter Plots: II

- $x_{t+k}$ versus $x_t$, $t = 1, \ldots, N - k$: lag $k$ scatter plot
- summarize scatter plots using linear model:
  
  $$x_{t+k} = \alpha_k + \beta_k x_t + \epsilon_{t,k}$$

  (not always reasonable: see Fig. 9)

- Pearson product moment correlation coefficient
  
  - let $y_1, \ldots, y_N$ & $z_1, \ldots, z_N$ be 2 collections of ordered values
  - let $\bar{y}$ & $\bar{z}$ be sample means (thus $\bar{y} \equiv \sum y_t / N$)
  - sample correlation coefficient:
    
    $$\hat{\rho} = \frac{\sum(y_t - \bar{y})(z_t - \bar{z})}{\left[\sum(y_t - \bar{y})^2 \sum(z_t - \bar{z})^2\right]^{1/2}},$$

  - measures strength of linearity ($-1 \leq \hat{\rho} \leq 1$)

Sample Autocorrelation Sequence

- let $\{y_t\} = \{x_{t+k} : t = 1, \ldots, N - k\}$
  
  and $\{z_t\} = \{x_t : t = 1, \ldots, N - k\}$

- for each lag $k$, plug these into
  
  $$\hat{\rho} = \frac{\sum(y_t - \bar{y})(z_t - \bar{z})}{\left[\sum(y_t - \bar{y})^2 \sum(z_t - \bar{z})^2\right]^{1/2}},$$

  and to get (after a little tweaking)
  
  $$\hat{\rho}_k \equiv \frac{\sum_{t=1}^{N-k}(x_{t+k} - \bar{x})(x_t - \bar{x})}{\sum_{t=1}^{N}(x_t - \bar{x})^2}$$

- $\hat{\rho}_k$, $k = 0, \ldots, N - 1$, called sample acs
- four examples (Figs. 6 and 7)
Sample ACS for Wind Speed Series

Sample ACS for Atomic Clock Series
Sample ACS for Willamette River Series

Sample ACS for Ocean Noise Series
Modeling of Time Series

- assume $x_t$ is realization of random variable $X_t$
- need to specify properties of $X_t$ (i.e., model $x_t$)
- simplifying assumptions (related to stationarity)
  - $\hat{\rho}_k$ estimates time-independent theoretical acs
    $\rho_k \equiv \text{cov}\{X_t, X_{t+k}\}/\sigma^2 \equiv E\{(X_t - \mu)(X_{t+k} - \mu)\}/\sigma^2,$
    where $\mu \equiv E\{X_t\}$ and $\sigma^2 \equiv E\{(X_t - \mu)^2\}$
  - $X_t$’s are multivariate Gaussian
- statistics of $X_t$’s completely known if $\mu$, $\sigma^2$ and $\rho_k$’s known
- critique of ‘time domain’ characterization ($\mu$, $\sigma^2$, $\rho_k$):
  - not easy to visualize $x_t$ from $\rho_k$’s
  - statistical properties of $\hat{\rho}_k$’s difficult to use

Frequency Domain Modeling: I

- idea: express $X_t$ in terms of cosines and sines (i.e., sinusoids)
- consider artificial time series $\cos(2\pi ft) \& \sin(2\pi ft)$, $t = 1, \ldots, 128$, where $f$ is the frequency of the sinusoid (and $1/f$ is the period)
- consider ten different frequencies (carefully chosen!):
  $$f = \frac{1}{128}, \frac{3}{128}, \ldots, \frac{17}{128}, \frac{19}{128}$$
- let $f_j = \frac{j}{128}$, where $j = 1, 3, \ldots, 19$
- in following twenty overheads, top plots show sinusoidal time series whose $t$th elements are
  $$\cos(2\pi f_1 t), \sin(2\pi f_1 t), \cos(2\pi f_3 t), \sin(2\pi f_3 t),$$
  $$\ldots, \cos(2\pi f_{19} t), \sin(2\pi f_{19} t)$$
Frequency Domain Modeling: II

- Bottom plots show cumulative sums of series:

  \[
  \begin{align*}
  &\cos(2\pi f_1 t) \\
  &\cos(2\pi f_1 t) + \sin(2\pi f_1 t) \\
  &\cos(2\pi f_1 t) + \sin(2\pi f_1 t) + \cos(2\pi f_3 t) \\
  &\cos(2\pi f_1 t) + \sin(2\pi f_1 t) + \cos(2\pi f_3 t) + \sin(2\pi f_3 t) \\
  &\vdots \\
  &\cos(2\pi f_1 t) + \sin(2\pi f_1 t) + \cdots + \cos(2\pi f_{19} t) \\
  &\cos(2\pi f_1 t) + \sin(2\pi f_1 t) + \cdots + \cos(2\pi f_{19} t) + \sin(2\pi f_{19} t)
  \end{align*}
  \]

Sinusoid and Sum of Sinusoids

\[f = 19/128\]

Sum of 20 sinusoids
Frequency Domain Modeling: III

- sum of all 20 sinusoids highly structured and nonrandom in appearance
- let’s repeat this exercise, but now multiply each sinusoid by a random amplitude $A$ (each sinusoid gets a different amplitude)
- $A$’s chosen from a standard Gaussian (normal) distribution (zero mean, unit variance)
Frequency Domain Modeling: IV

- generalize to following simple model for $X_t$:

$$X_t = \mu + \sum_{j=1}^{N/2} [A_j \cos (2\pi f_j t) + B_j \sin (2\pi f_j t)]$$

- holds for $t = 1, 2, \ldots, N$, where $N$ is even
- $f_j \equiv j/N$ fixed frequencies (cycles/unit time)
  (called Fourier or standard frequencies)
- $A_j$’s and $B_j$’s are random variables:
  * $E\{A_j\} = E\{B_j\} = 0$
  * $\text{var} \{A_j\} = \text{var} \{B_j\} = \sigma_j^2$ (now allowed to depend on $j$)
  * $\text{cov} \{A_j, A_k\} = \text{cov} \{B_j, B_k\} = 0$ for $j \neq k$
  * $\text{cov} \{A_j, B_k\} = 0$ for all $j, k$

The Spectrum: I

- properties of simple model (Exercise [1.1]):
  - $E\{X_t\} = \mu$
  - $\sigma_j^2$’s decompose population variance:

$$\sigma^2 = E\{(X_t - \mu)^2\} = \sum_{j=1}^{N/2} \sigma_j^2$$

- $\sigma_j^2$’s determine acs:

$$\rho_k = \frac{\sum_{j=1}^{N/2} \sigma_j^2 \cos (2\pi f_j k)}{\sum_{j=1}^{N/2} \sigma_j^2}$$

- define spectrum as $S_j \equiv \sigma_j^2$, $1 \leq j \leq N/2$
The Spectrum: II

- fundamental relationship:
  \[ \sum_{j=1}^{N/2} S_j = \sigma^2 \]
  - decomposes \( \sigma^2 \) into components related to \( f_j \)
  - \( S_j \)'s equivalent to acs and \( \sigma^2 \) (Exercise [1.5])
- easy to simulate \( x_t \)'s from simple model
- four examples of
  - spectra versus \( f_j \)
  - acs's versus \( k \)
  - \( x_t \)'s versus \( t \)

\[ \text{Theoretical Spectrum for Wind Speed Series} \]
Theoretical and Sample ACSs for Wind Speed

Actual and Simulated Wind Speed Series
Theoretical Spectrum for Atomic Clock Series

Theoretical and Sample ACSs for Atomic Clock
Nonparametric Estimation of $S_j$: I

- problem: estimate spectrum $S_j$ from $X_1, \ldots, X_N$
- mine out $A_j$’s & $B_j$’s since $S_j = \text{var} \{A_j\} = \text{var} \{B_j\}$
- could use linear algebra ($N$ knowns and $N$ unknowns)
- can get $A_j$’s via discrete Fourier cosine transform since

$$\sum_{t=1}^{N} X_t \cos(2\pi f_j t) = \frac{NA_j}{2}$$

- yields (for $1 \leq j < N/2$): $A_j = \frac{2}{N} \sum_{t=1}^{N} X_t \cos(2\pi f_j t)$
Nonparametric Estimation of $S_j$: II

- $B_j$’s from sine transform: $B_j = \frac{2}{N} \sum_{t=1}^{N} X_t \sin (2\pi f_j t)$

- since $S_j = \text{var} \{ A_j \} = \text{var} \{ B_j \}$, can estimate $S_j$ using

\[
\hat{S}_j \equiv \frac{A_j^2 + B_j^2}{2} = \frac{2}{N^2} \left[ \left( \frac{\sum_{t=1}^{N} X_t \cos (2\pi f_j t)}{N} \right)^2 + \left( \frac{\sum_{t=1}^{N} X_t \sin (2\pi f_j t)}{N} \right)^2 \right]
\]

- examples: Figs. 20 and 21
Theoretical/Estimated Spectra for Atomic Clock

Theoretical/Estimated Spectra for Willamette River
Nonparametric Estimation of $S_j$: III

- points about $\hat{S}_j$
  - uncorrelatedness of $A_j$'s and $B_j$'s implies $\hat{S}_j$'s approximately uncorrelated (exact under Gaussian assumption)
  - easy to test hypothesis using $\hat{S}_j$'s (difficult for sample acs)
  - $\hat{S}_j$ is ‘2 degrees of freedom’ estimate; if $S_j$’s slowly varying, can average $\hat{S}_j$’s locally
Parametric Estimation of $S_j$: I

- assume $S_j$’s depend on small number of parameters
- simple model:
  \[
  S_j(\alpha, \beta) = \frac{\beta}{1 + \alpha^2 - 2\alpha \cos(2\pi f_j)}
  \]
  (related to first-order autoregressive process)
- estimate $S_j$’s by estimating $\alpha, \beta$:
  \[
  \hat{S}_j(\hat{\alpha}, \hat{\beta}) = \frac{\hat{\beta}}{1 + \hat{\alpha}^2 - 2\hat{\alpha} \cos(2\pi f_j)}
  \]

Parametric Estimation of $S_j$: II

- can show that $\rho_1 \approx \alpha$, so let $\hat{\alpha} = \hat{\rho}_1$
- requiring
  \[
  \frac{N}{2} \sum_{j=1}^{N/2} \hat{S}_j(\hat{\alpha}, \hat{\beta}) = \frac{1}{N} \sum_{t=1}^{N} (X_t - \bar{X})^2 \equiv \hat{\sigma}^2
  \]
  yields estimator
  \[
  \hat{\beta} = \hat{\sigma}^2 \left( \sum_{j=1}^{N/2} \frac{1}{1 + \hat{\alpha}^2 - 2\hat{\alpha} \cos(2\pi f_j)} \right)^{-1}
  \]
- examples: ‘theoretical’ spectra for wind speed, atomic clock and ocean noise (doesn’t work well Willamette River series, which points out need to be careful about parameterization)
Parametric/Nonparametric Estimated Spectra for Wind Speed

Atomic Clock

Parametric/Nonparametric Estimated Spectra for Wind Speed
Parametric/Nonparametric Estimated Spectra for Ocean Noise

'Industrial Strength' Theory: I

- simple model not adequate in practice
  - frequencies in model tied to sample size \( N \)
  - time series treated as if it were 'circular'; i.e.,
    \[
    X_k, X_{k+1}, \ldots, X_{N-1}, X_N, X_1, X_2, \ldots, X_{k-1}
    \]
    has same spectrum as \( X_1, X_2, \ldots, X_N \).
- assume stationarity, which means that
  \[
  E\{X_t\} = \mu, \quad \text{var} \{X_t\} = \sigma^2 \quad \text{and} \quad \text{cov} \{X_t, X_{t+k}\} = \rho_k \sigma^2,
  \]
‘Industrial Strength’ Theory: II

• under stationarity, simple model extends to become

\[ X_t = \mu + \int_{-1/2}^{1/2} e^{i2\pi ft} dZ(f) \]

\[ \approx \mu + \sum_f [A(f) \cos(2\pi ft) + B(f) \sin(2\pi ft)] , \]

where \( dZ(f) \) yields \( A(f) \) and \( B(f) \), and we now use

\[ e^{i2\pi ft} \equiv \cos(2\pi ft) + i \sin(2\pi ft), \quad i \equiv \sqrt{-1} \]

• analogous to simple model, we use

\[ \text{var} \{dZ(f)\} = S(f) \text{ df} \]

to define a spectral density function \( S(f) \)

I–53

‘Industrial Strength’ Theory: III

• fundamental relationship now becomes

\[ \int_{-1/2}^{1/2} S(f) \text{ df} = \sigma^2 \]

• \( S(f) \) and \( \rho_k \sigma^2 \) related via

\[ \rho_k \sigma^2 = \int_{-1/2}^{1/2} S(f) e^{i2\pi fk} \text{ df} \quad \text{and} \quad S(f) = \sigma^2 \sum_{k=-\infty}^{\infty} \rho_k e^{-i2\pi fk} \]

• basic estimator of \( S(f) \) is periodogram:

\[ \hat{S}^{(p)}(f) \equiv \frac{1}{N} \left| \sum_{t=1}^{N} (X_t - \overline{X}) e^{-i2\pi ft} \right|^2 , \quad \text{where} \quad \overline{X} \equiv \frac{1}{N} \sum_{t=1}^{N} X_t \]

I–54
‘Industrial Strength’ Theory: IV

- ideally it would be nice if
  1. $E\{\hat{S}(p)(f)\} = S(f)$
  2. $\text{var}\{\hat{S}(p)(f)\} \to 0$ as $N \to \infty$

but, alas,

1. periodogram can be badly biased for finite $N$ (can correct using data tapers)
2. $\text{var}\{\hat{S}(p)(f)\} = S^2(f)$ as $N \to \infty$ if $0 < f < \frac{1}{2}$ (can correct using smoothing windows)

Uses of Spectral Analysis

- analysis of variance technique for time series
- some uses
  - testing theories (e.g., wind data)
  - exploratory data analysis (e.g., rainfall data)
  - discriminating data (e.g., neonates)
  - diagnostic tests (e.g., ARIMA modeling)
  - assessing predictability (e.g., atomic clocks)
- applications
  - tout le monde!