Introduction to Spectral Analysis

Don Percival

Applied Physics Lab, University of Washington

overheads available at

`http://www.staff.washington.edu/dbp/talks.html`

What is Spectral Analysis?

- one of the most widely used (and lucrative!) methods in data analysis
- can be regarded as
  - analysis of variance of time series using sinusoids
  - sinusoids + statistics
  - Fourier theory + statistics
- today's lecture: introduction to spectral analysis
  - notion of a 'time' series
  - $0.25 introduction to time series analysis
    - basics of 'time domain' analysis
    - subject of Stat 519
  - notion of the spectrum
  - methods for estimating the spectrum
    - nonparametric
    - parametric
  - concluding comments
  - Stat/EE 520 has (lots!) more details

Time Series & Time Series Analysis

- what is a time series?
  - 'one damned thing after another' (R. A. Fisher?)
  - x_t, $t = 1, \ldots, N$
  - four examples (Figures 2 and 3)
- goal of time series analysis:
  - quantify characteristics of time series
- univariate statistics, e.g., sample mean & variance

$$\bar{x} \equiv \frac{1}{N} \sum_{t=1}^{N} x_t \text{ and } \hat{\sigma}^2 \equiv \frac{1}{N} \sum_{t=1}^{N} (x_t - \bar{x})^2,$$

inadequate to say how x_t and x_{t+k} are related

Lagged Scatter Plots

- bivariate distribution of separated pairs
- x_{t+1} versus x_t, $t = 1, \ldots, N-1$: lag 1 scatter plot
- four examples (Figure 4)
- 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 Figure 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
  - sample correlation coefficient:

$$\hat{\rho} = \frac{\Sigma(y_t - \bar{y})(z_t - \bar{z})}{\left[\Sigma(y_t - \bar{y})^2 \, \Sigma(z_t - \bar{z})^2\right]^{1/2}},$$

  - 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

$$\hat{\rho} = \frac{\Sigma(y_t - \bar{y})(z_t - \bar{z})}{\left[\Sigma(y_t - \bar{y})^2 \, \Sigma(z_t - \bar{z})^2\right]^{1/2}},$$

  and fudge things a bit to get

$$\hat{\rho}_k \equiv \frac{\Sigma_{t=1}^{N-k}(x_{t+k} - \bar{x})(x_t - \bar{x})}{\Sigma_{t=1}^{N}(x_t - \bar{x})^2}$$

- $\hat{\rho}_k$, $k = 0, \ldots, N-1$, called sample acs
- four examples (Figures 6 and 7)

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

    $\rho_k \equiv \text{cov}\,\{X_t, X_{t+k}\}/\sigma^2 \equiv E\{(X_t - \mu)(X_{t+k} - \mu)\}/\sigma^2,$

    where
    * $\mu \equiv E\{X_t\}$ (note: does not depend on t)
    * $\sigma^2 = E\{(X_t - \mu)^2\}$ (does not depend on t)
  - X_t's are multivariate Gaussian
  - statistics of X_t's completely determined if we know μ, σ^2 and ρ_k's
- critique of 'time domain' characterization (μ, σ^2, ρ_k):
  - not easy to visualize x_t from ρ_k's
  - statistical properties of $\hat{\rho}_k$'s difficult to use

Frequency Domain Modeling: I

- based on idea of expressing X_t in terms of sinusoids
- top five rows of Figure A1 show $\cos(2\pi f t)$ for
  $$t = 1, \ldots, 128 \;\&\; f = \tfrac{1}{128}, \tfrac{1}{8}, \tfrac{1}{4}, \tfrac{3}{8}, \tfrac{1}{2},$$
  where f is frequency of sinusoid ($1/f$ is period)
- bottom row shows addition of five sinusoids
  - highly structured and nonrandom
- Figure A2 shows $\cos(2\pi f t + \phi)$ with ϕ chosen randomly (one for each f)
  - rattier looking, but still highly structured
- Figure A3 shows additions of 64 sinusoids with frequencies $\frac{1}{128}$, $\frac{2}{128}$ $\ldots$, $\frac{63}{128}$, $\frac{64}{128}$ & random phases
  - very ratty looking, with no apparent structure
- note: $\cos(2\pi f t + \phi) = A\cos(2\pi f t) + B\sin(2\pi f t)$, where $A = \cos(\phi)$ and $B = -\sin(\phi)$
  - $E\{A\} = E\{B\} = 0$
  - $\operatorname{var}\{A\} = \operatorname{var}\{B\} = \frac{1}{2}$
  - $\operatorname{cov}\{A, B\} = 0$, i.e., uncorrelated (?!)

Frequency Domain Modeling: II

- generalize to following simple model for X_t:

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

  - 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:
    - $E\{A_j\} = E\{B_j\} = 0$
    - $\operatorname{var}\{A_j\} = \operatorname{var}\{B_j\} = \sigma_j^2$
    - $\operatorname{cov}\{A_j, A_k\} = \operatorname{cov}\{B_j, B_k\} = 0$ for $j \neq k$
    - $\operatorname{cov}\{A_j, B_k\} = 0$ for all j, k
  - note: σ_j^2 now allowed to depend on j

The Spectrum

- properties of simple model:
  - $E\{X_t\} = \mu$
  - σ_j^2's decompose population variance:

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

  - σ_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 $S_j \equiv \sigma_j^2$, $1 \leq j \leq N/2$
- fundamental relationship:

$$\sum_{j=1}^{N/2} S_j = \sigma^2$$

  - decomposes σ^2 into components related to f_j
  - S_j's equivalent to acs and σ^2
- easy to simulate x_t's from simple model
- examples of spectra (in dB), acs's and x_t's (Figures 12 to 17)

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:

$$\begin{aligned}
\sum_{t=1}^{N} X_t \cos(2\pi f_j t) &= \mu \sum_{t=1}^{N} \cos(2\pi f_j t) \\
&\quad + \sum_{t=1}^{N} \sum_{k=1}^{N/2} A_k \cos(2\pi f_k t) \cos(2\pi f_j t) \\
&\quad + \sum_{t=1}^{N} \sum_{k=1}^{N/2} B_k \sin(2\pi f_k t) \cos(2\pi f_j t) \\
&= \sum_{k=1}^{N/2} A_k \sum_{t=1}^{N} \cos(2\pi f_k t) \cos(2\pi f_j t) \\
&\quad + \sum_{k=1}^{N/2} B_k \sum_{t=1}^{N} \sin(2\pi f_k t) \cos(2\pi f_j t) \\
&= \frac{N A_j}{2}
\end{aligned}$$

- yields (for $1 \leq j < N/2$): $A_j = \dfrac{2}{N} \sum_{t=1}^{N} X_t \cos(2\pi f_j t)$
- B_j's from sine transform: $B_j = \dfrac{2}{N} \sum_{t=1}^{N} X_t \sin(2\pi f_j t)$

Nonparametric Estimation of S_j: II

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

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

- examples: Figures 20 and 21
- 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
  - $\log(\hat{S}_j)$ stabilizes variance (rationale for dB's)

Parametric Estimation of S_j

- 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 α, β:
$$\hat{S}_j(\hat{\alpha}, \hat{\beta}) = \frac{\hat{\beta}}{1 + \hat{\alpha}^2 - 2\hat{\alpha} \cos{(2\pi f_j)}}$$
  - can show that $\rho_1 \approx \alpha$, so let $\hat{\alpha} = \hat{\rho}_1$
  - requiring
$$\sum_{j=1}^{N/2} \hat{S}_j(\hat{\alpha}, \hat{\beta}) = \frac{1}{N} \sum_{t=1}^{N} (X_t - \bar{X})^2 \equiv \hat{\sigma}^2$$
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: thicks curves on Figures 20 and 21
- need to be careful about parameterization
(model here poor for Willamette River spectrum)

‘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 $X_1, X_2, \ldots, X_N$.
- under assumption of stationarity, i.e.,

  $$E\{X_t\} = \mu, \ \ \mathrm{var}\,\{X_t\} = \sigma^2 \ \text{ and } \ \mathrm{cov}\,\{X_t, X_{t+k}\} = \rho_k \sigma^2$$

  simple model extends to become

$$X_t = \mu + \int_{-1/2}^{1/2} e^{i2\pi ft}\, dZ(f) \approx \sum_f \left[A(f)\cos(2\pi ft) + B(f)\sin(2\pi ft)\right],$$

  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

  $$\mathrm{var}\,\{dZ(f)\} = S(f)\,df$$

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

‘Industrial Strength’ Theory: II

- fundamental relationship now becomes

$$\int_{-1/2}^{1/2} S(f)\,df = \sigma^2$$

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

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

- basic estimator of $S(f)$ is periodogram:

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

- 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!