Data Analysis for Earth, Marine, and Environmental Sciences

Jonathan & Eitan Lees

2024-12-02

Table of contents

Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

1 Introduction

This is a book created from markdown and executable code.

See Knuth (1984) for additional discussion of literate programming.

Part I

Time Series

This section is about time series analysis.

2 Fourier Analysis

This section is an introduction to Fourier Analysis. We will cover a variety of topics including

- Complex Numbers Review
- Series Expansions: exp, cosine, sine Euler's Formulae
- Definition of Fourier Transform (Continuous) Fourier Transform Pairs Amplitude and Phase
- Frequency, Period, Sampling
- Nyquist Frequency
- Convolution vs. Correlation Periodogram
- Leakage and Tapering

2.1 Fourier Basic Idea

How would you describe this signal?:

```
dt = 1/100

t = seq(from=0, by=dt, length=200);

y1 = 10*sin(2*pi*.2*t)
y2 = .4*sin(2*pi* 3*t);
y3 = y1+ y2

par(mai=c(.5, .5, .1,.1) )

plot(t, y3, type='l', xlab='', ylab='')
```

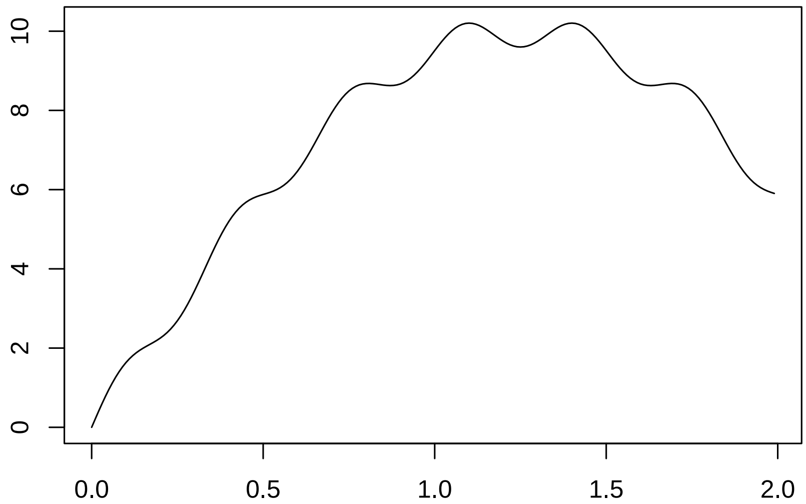


Figure 2.1: A single time series

The signal can be represented as a sum of different sinusoids:

- Signal S1: .2 Hz amplitude 10
- Signal S2: 3 Hz amplitude .4
- Signal S3 = S1 + S2

```
par(mfrow=c(3,1) )
par(mai=c(.5, .5, .1,.1) )
plot(t, y1, ylim=range(y3) , type='l', xlab='', ylab='')
plot(t, y2, ylim=range(y3), type='l', xlab='', ylab='')
plot(t, y3, ylim=range(y3), type='l', xlab='', ylab='')
```

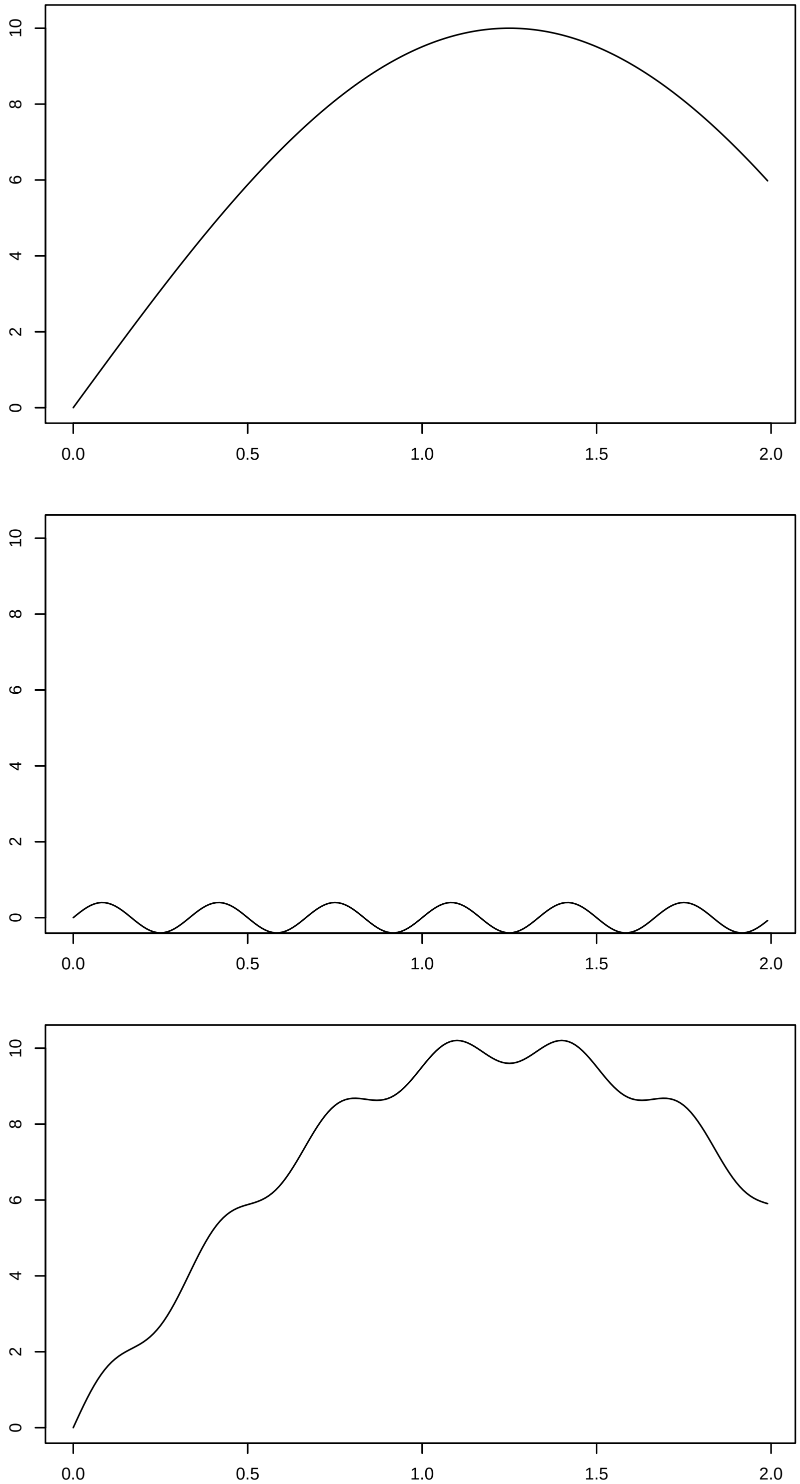


Figure 2.2: The same signal can be represented as a sum of sinusoids.

One could convey all the information with 6 numbers:

- dt, length
- 10, .2
- .4, 3

2.2 Discrete Sampling

```
library(RSEIS)

dt = 0.001
omega = 4
A = 2
phi0 = 20*pi/180

x = seq(from=0, to=5, by=dt)
y = A*sin(omega*x - phi0)

aa <- which(peaks(y, span=3))

ay1 = y[aa[1]]+.1*A
ay2 = y[aa[2]]+.1*A

plot(range(x), range(y), type='n', ann=FALSE, axes=FALSE)
lines(x,y)
axis(1)
axis(2)

DT =0.2
g = seq(from=min(x), to=max(x), by=DT)
gy = A*sin(omega*g - phi0)

points(g,gy)
segments(g, rep(0, times=length(g)), g, gy, col=grey(0.85))
abline(h=0)


segments(x[aa[1]], ay1,x[aa[2]], ay2, lwd=2, col='red', xpd=TRUE)
```

```
segments(x[aa[1]], ay1,x[aa[1]],y[aa[1]] , col='red', xpd=TRUE)
segments(x[aa[2]], ay2,x[aa[2]],y[aa[2]] , col='red', xpd=TRUE)

segments(x[aa[3]], 0,x[aa[3]],y[aa[3]] , col='blue', lwd=2, xpd=TRUE)
```

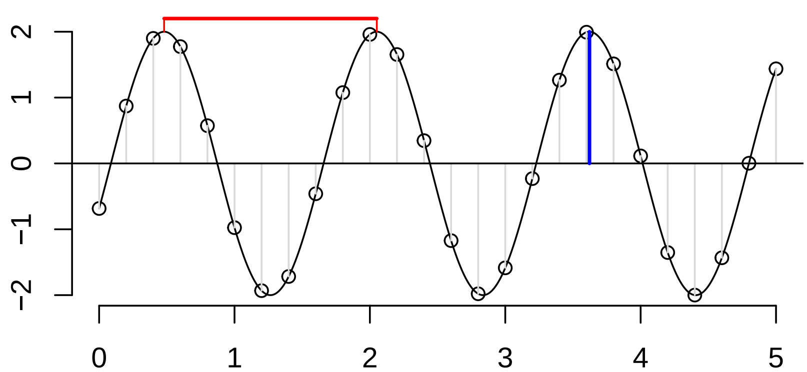


Figure 2.3: A wave is sampled at discrete points in time.

- Red: Period of Sine Wave (s)
- Blue: Amplitude of Sine Wave
- δt : sampling rate (s)

2.3 Cycles, Phase and Frequency

The rotating pen height in Figure 2.4 represents the signal.

$$Y_i = A \sin (2\pi x_i/X + \phi)$$

$$\alpha_i = (2\pi x_i/X + \phi)$$

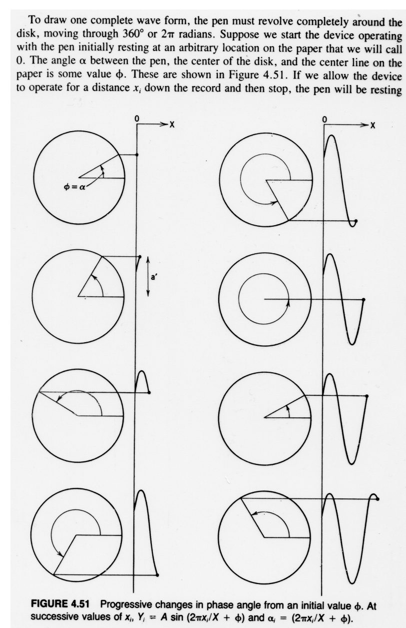


Figure 2.4: Cycles and Sines

2.4 Time Series: Basics

- Signal Characteristics:
  - period = T/cycle
  - frequency $f = 1/T$ cycles/s
- Sampling

- sample rate $=\Delta t$
- Sampling Frequency $= \frac{1}{\Delta t} = f_{sampling}$

Units:
if $y = \cos(\theta) = \cos(\omega t)$

- θ is in units of radians $= 2\pi f t$
- there are 2π radians per cycle
- we define the angular frequency $\omega = 2\pi f$ radians/sec
- time t is defined as $t = i \cdot \Delta t$ where i is the sample
- $T = \sum i \cdot \Delta t$ is the total time

$$y_k = A\cos(\omega t - \phi)$$

where ϕ is the phase and ω is the frequency.

$$\begin{aligned} y_k &= A\cos(\omega t - \phi) \\ &= A\cos(\omega t)\cos\phi + A\sin(\omega t)\sin\phi \\ &= \alpha_k \cos(\omega t) + \beta_k \sin(\omega t) \end{aligned}$$

Trig Identities

$$\begin{aligned} \sin(u \pm v) &= \sin u \cos v \pm \cos u \sin v \\ \cos(u \pm v) &= \cos u \cos v \mp \sin u \sin v \end{aligned}$$

2.5 Trig Functions

Consider the Taylor series expansions:

$$\begin{aligned} e^x &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \\ \sin(x) &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots \\ \cos(x) &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots \end{aligned}$$

Plug in ix in the formula for e^x get:

$$e^{ix} = 1 + ix + \frac{(ix)^2}{2!} + \frac{(ix)^3}{3!} + \frac{(ix)^4}{4!} + \frac{(ix)^5}{5!} + \frac{(ix)^6}{6!} + \frac{(ix)^7}{7!} + \cdots$$

Expanding out the complex numbers gives the trigonometric functions.

$$e^{ix} = \cos(x) + i\sin(x)$$

2.6 Euler's Formula

$$e^{i\theta} = \cos(\theta) + i\sin(\theta)$$

And

$$e^{-i\theta} = \cos(\theta) - i\sin(\theta)$$

Add these together:

$$e^{i\theta} + e^{-i\theta} = 2\cos(\theta)$$

or:

$$\cos(\theta) = \frac{e^{i\theta} + e^{-i\theta}}{2}$$

Similarly, by subtracting:

$$\sin(\theta) = \frac{e^{i\theta} - e^{-i\theta}}{2i}$$

2.7 Fourier Analysis

- The Fourier Transform (FT) is a series of complex numbers $[a, b] = a + ib$.
- The real and imaginary parts of the FT can be combined to extract different information from the FT.
- The Amplitude spectrum is the modulus of the complex numbers:

  $$A_i = \sqrt{a_i^2 + b_i^2}$$

- The phase spectrum is the phase angle:

  $$\phi_i = \tan^{-1}\left(\frac{b_i}{a_i}\right)$$

2.8 Links to animations

- Geometric Fourier Transform Animation by Michael Borcherds
- What is the Fourier Transform? by 3Blue1Brown

2.9 Fourier Series

$$
\begin{aligned}
y_k &= A\cos(\omega t - \phi) \\
&= A\cos\omega t\cos\phi + A\sin\omega t\sin\phi \\
&= \alpha_k\cos(\omega t) + \beta_k\sin(\omega t)
\end{aligned}
$$

The Fourier Coefficients are α_k, β_k

This leads to Fourier's Theorem:

$$Y = \sum_{k=0}^{\infty} A_k\cos(k\theta + \phi_k)$$

$$
\begin{aligned}
\beta_k &= \frac{2}{k}\sum_{j=0}^{n-1} Y_j \sin\left(\frac{2\pi jk}{n}\right) \\
\alpha_k &= \frac{2}{k}\sum_{j=0}^{n-1} Y_j \cos\left(\frac{2\pi jk}{n}\right)
\end{aligned}
$$

The Zero-th value of the FT is the Mean value of the time series:

$$\alpha_0 = \frac{1}{n}\sum_{j=0}^{n-1} Y_j$$

This is usually called the DC or "direct current".

Given the definition of the Fourier Series above, the spectrum is defined as:

$$
\begin{aligned}
A_i &= \sqrt{a_i^2 + b_i^2} \\
\phi_i &= \tan^{-1}\left(\frac{b_i}{a_i}\right)
\end{aligned}
$$

Figure 2.5: FFT Explained

2.10 Fourier Analysis

- period = T/cycle
- sample rate $=\Delta t$
- frequency $f = 1/T$ cycles/s
- Sampling Frequency $= \frac{1}{\Delta t} = f_{sampling}$

if $y = \cos(\theta) = \cos(\omega t)$

- θ is in units of radians $= 2\pi f t$
- there are 2π radians per cycle
- we define the angular frequency $\omega = 2\pi f$ radians/sec
- time t is defined as $t = \frac{(i \cdot \Delta t)}{T}$ where T is the total time

2.11 Fourier Transform

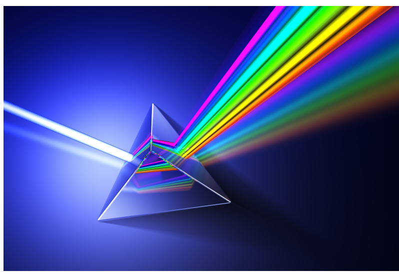
The Fourier transform of a function $f(x)$ is a complex valued function $F(\omega)$

Fourier Transform

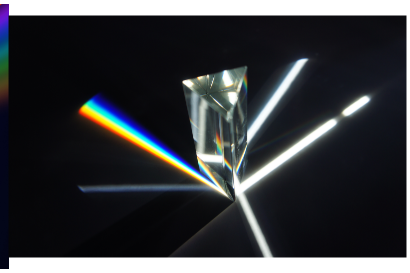
$$FT(f) = F(\omega) = \int_{-\infty}^{\infty} f(x)e^{-ix\omega}dx$$

Remembering that $\omega = 2\pi f$

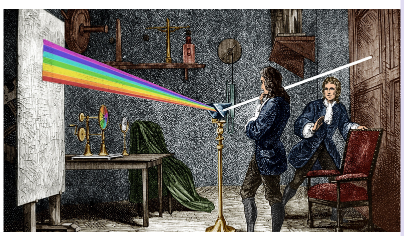
2.12 Fourier Transform: Prism



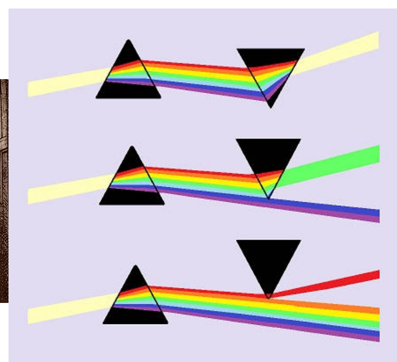
(a) Illustration of a prism



(b) A real prism



(c) Newton's light experiements



(d) Newtons critical insight

Figure 2.6: The FFT is a timeseries prism.

Think of the Fourier Transform like a prism: the input has numerous signals all combined: the FT separates the signals into sinusoidal elements and assigns a ‘power’, or level, to each component.

2.13 Fourier Analysis: R

Suppose you have a time series, g that has n samples:

- In **R** you can get the Fourier transform by using the function fft (Fast Fourier Transform)
- fft works fastest when the number of samples is a power of 2
- if n is not a power of 2, can zero-pad to closest power
- the fft function returns a complex valued vector n-samples long
- it is a good idea to remove the mean from the signal before fft
- the fft is symmetric: usually we display only half
- Use functions Mod, and Arg to extract the Modulus and Phase Angle
- Parseval's theorem:

$$\sum^{n} \text{Abs}(fft)^2 = \sum^{n} \text{Abs}(g)^2$$

2.14 Sampling and Aliasing

- In nearly all cases in data analysis in the earth sciences we sample the data at discrete intervals.
- This means that signals are never continuous.
- We can think of this process as multiplying the underlying continuous (natural) signal by a comb function and a boxcar function.
- the boxcar function is applied because our observations have a finite time interval.

The Nyquist theorem states that we must sample an underlying signal at least twice per cycle in order to reconstruct a particular frequency. Or, if we sample at a rate of Δt, then

$$f_{Nyquist} = \frac{1}{2\Delta t}$$

is the maximum frequency we can extract without aliasing.

2.15 Fourier Analysis

- Amplitude spectrum:

$$A = \sqrt{a^2 + b^2}$$

- Phase spectrum:

$$\phi = \tan^{-1}\left(\frac{b}{a}\right)$$

We can think of the amplitude spectrum is offering information on the statistical properties of the underlying time series: How much variance of the original signal is accounted for in each Fourier component?

This is the underlying concept of the *Power Spectrum*.

2.16 Convolution and Correlation

Convolution and Correlation of two time series are related:

Cross Correlation

$$(f \star g)(t) \equiv \int_{-\infty}^{\infty} f^*(\tau) g(t+\tau) d\tau$$

To get the correlation: shift, multiply, sum

Convolution (Time reversed correlation)

$$(f \circledast g)(t) \equiv \int_{-\infty}^{\infty} f(\tau) g(t-\tau) d\tau$$

2.17 Shift Theorem

Shifting the time series multiplies the FT by a complex exponential.

Theorem

$$FT(g(x-a)) = e^{-i\omega a}G(\omega)$$

Start with the definition of the FT: $FT(g) = \int g e^{-i\omega t} dt$

Proof

$$\int_{-\infty}^{\infty} f(x-a)e^{-i2\pi xs}dx$$

Substitute $u = x - a$, so that $du = dx$ and $x = u + a$:

$$\int_{-\infty}^{\infty} f(u)e^{-i2\pi(u+a)s}du = \int_{-\infty}^{\infty} f(u)e^{-i2\pi us}e^{-i2\pi as}du = e^{-i2\pi as}F(s)$$

2.18 Convolution Theorem

Prove convolution theorem:

$$FT\left[\int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau\right] = F(\omega)G(\omega)$$

Or, convolution in the time domain is multiplication in the frequency domain.

Proof of Convolution Theorem

$$\begin{aligned}
FT\left[\int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau\right] &= \\
\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\tau)g(t-\tau)e^{-i\omega t}d\tau dt &= \\
\int_{-\infty}^{\infty} f(\tau)\int_{-\infty}^{\infty} g(t-\tau)e^{-i\omega t}dtd\tau &= \\
\int_{-\infty}^{\infty} f(\tau)e^{-i\omega\tau}G(\omega)d\tau &= F(\omega)G(\omega) \quad \square
\end{aligned}$$

Multiplication in the time domain is convolution in the frequency domain

$$g(t) \times f(t) \Leftrightarrow F(\omega) \circledast G(\omega)$$

Multiplication in the frequency domain is convolution in the time domain

$$F(\omega) \times G(\omega) \Leftrightarrow g(t) \circledast f(t)$$

2.19 Convolution

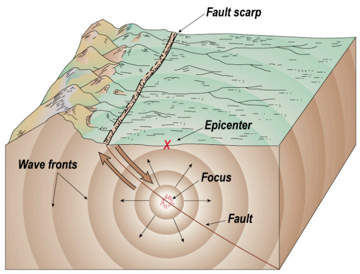
- Any discrete measurement of a continuous process
- Seismogram
- Climate Cycles
- Filtering
- Convolution is the way we describe the interaction of signal processes

Example: Seismic Data

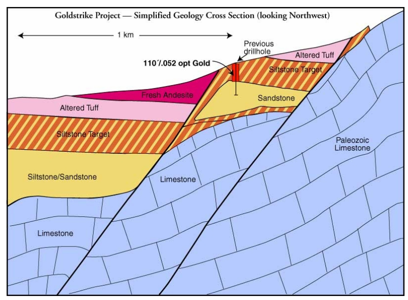
- Source
- Earth Structure
- Instrument
- An observed signal can be modeled as a convolution of these processes

$$Source \bigotimes Earth \bigotimes Instrument = Signal$$

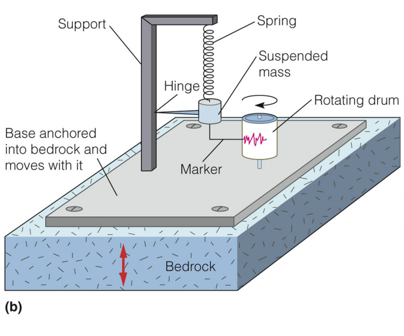
```
library(RSEIS)

dt = 0.01
freq = 16
nlen = 35

G = genrick(freq, dt, nlen)

tee = seq(from=0, by=dt, length=length(G))

x = sin(2*pi*freq*tee)*exp(-10*tee)
```



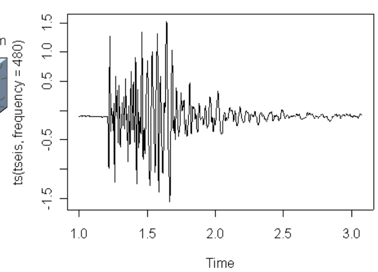
(a) An earthquake occures in the earths crust.



(b) The composition and geometry of the fault effect the propagation.



(c) A seismometer records an event at the surface.



(d) Finally we have a signal to analyze.

Figure 2.7: A signal can be thought of as a convolution of source, earth, and instrument.

```
rx = rev(x)

mt1 = min(tee)
mt2 = max(tee)

x1 = min(x)
x2 = max(x)

rx1 = min(rx)
rx2 = max(rx)


spiks = rep(0, 50)
spiks[22] = 1
spiks[34] = 1

s1 = min(spiks)
s2 = max(spiks)

timesp = seq(from=0, by=dt, length=length(spiks))

st1 = min(timesp )
st2 = max(timesp )

c2 = convolve(x, spiks, type = c("open"))
ct = seq(from=0, by=dt, length=length(c2))
c21 = min(c2 )
c22 = max(c2 )

ct1 = min(ct)
ct2 = max(ct)

####
library(RPMG)

par(mai=c(.0, .0, .0,.0) )
plot(c(0,1), c(0,1.2), type='n', ann=FALSE, axes=FALSE)

lines(RESCALE(tee, 0, .4, mt1, mt2), RESCALE(x, 0.8, 1.0, x1, x2), col="blue")
text(0, 1.1, "Instrument Response", pos=4)
```

```
lines(RESCALE(timesp, .6, .9, st1, st2), RESCALE(spiks, 0.8, 1.0, s1, s2), col="blue")

text(0.6, 1.1, "Earth Response", pos=4)

lines(RESCALE(tee, 0, .4, mt1, mt2), RESCALE(rx, 0.5, .7, rx1, rx2), col="red")
text(0, 0.7, "Reversed Instrument Response", pos=4)

lines(RESCALE(timesp, .6, .9, st1, st2), RESCALE(spiks, 0.5, 0.7, s1, s2), col="blue")
arrows(.45, .6, .55, .6)

lines(RESCALE(ct, .2, .8, ct1, ct2), RESCALE(c2, 0.1, 0.3, c21, c22), col="purple")

text(.2 , 0.35, "Convolved Output", pos=4)
```

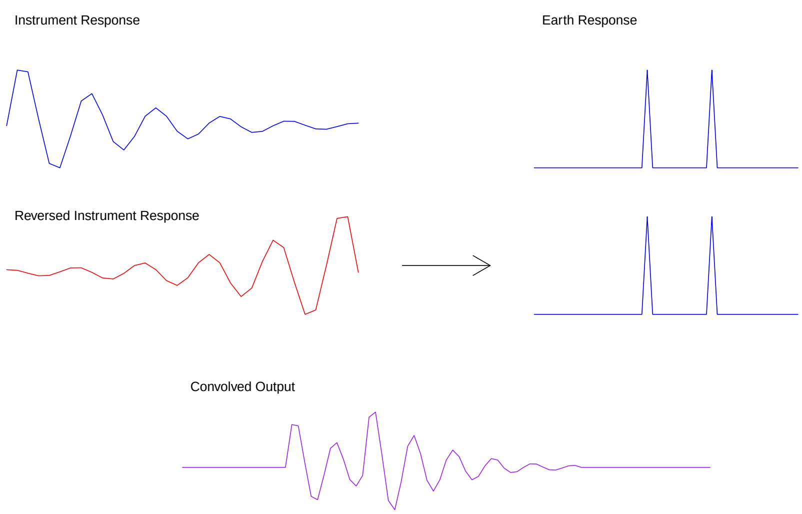


Figure 2.8: An example of convolution

Figure 2.9: Convolution Explained

2.20 Convolution: Thermometer Reading

- If you are measuring a temperature and the ambient temperature suddenly drops
- What do you observe?
- Observe = Temp $\circledast$ Thermometer
- The observation is the step function of the temperature convolved with the response function of the thermometer

2.21 Convolution

- Convolution is correlation with one of the time series reversed
- A $\circledast$ B = correlate A(t) with B(-t)
- or: flip one time series and correlate

Convolution is the way processes interact in the earth. This is a model, of course, but it seems to work.

$$C(t) = \int_{-\infty}^{\infty} A(t)B(-t)dt$$

Convolution is the cross correlation of one time series with the time-reversed version of another time series.

2.22 Periodogram

Recall the definition of the Autocorrelation:

$$Auto(\tau) = \frac{E\left[(X_t - \mu)(X_{t+\tau} - \mu)\right]}{\sigma^2}$$

$$FT\left[\int_{-\infty}^{\infty} f(\tau)g(t+\tau)d\tau\right] = F(\omega)G^*(\omega)$$

Let $g = f$ in the convolution theorem, get Fourier Transform *Autocorrelation*:

$$FT\left[\int_{-\infty}^{\infty} f(\tau)f(t+\tau)d\tau\right] = F(\omega)F^*(\omega) = |F(\omega)|^2$$

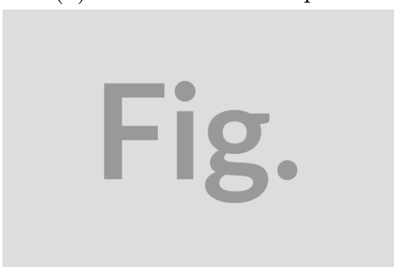
This is commonly called the *periodogram.* It is a simple measure of the variance of each fourier component (sinusoid) represented in the signal.

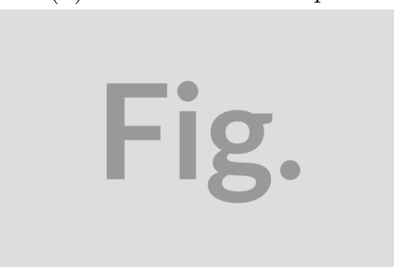
> ⚠ Warning
>
> The periodogram is not a good estimator of spectrum.

- The periodogram is not a consistent estimator of the true underlying spectrum
- Adding more data increases frequency resolution, but does not reduce variance
- Must devise smoothing method to get around this problem
- Smooth the periodogram
- Average multiple spectra from multiple realizations of the time series (welch's method)

- We usually scale the power spectra using one of several methods:
  - N (periodogram)
  - Var(Y)

(a) DFT with 64 samples (b) DFT with 128 samples





(c) DFT with 256 samples (d) DFT with 1024 samples

Figure 2.10: Periodogram example

Figure 2.11: Raw Spectrum

(a) frequency bandwidth = 0.00113

(b) frequency bandwidth = 0.00298

(c) frequency bandwidth = 0.00564

(d) frequency bandwidth = 0.00851

Figure 2.12: Smoothed Periodogram example

  - Std(Y)

- Sometimes it is useful to remove all scales from the spectrum: plot as decibels
- A decibel is the log of the ratio of amplitudes
- $dB = 10 \log(A/A_0)$
- If the power spectrum is needed, use
- $dB = 20 \log(A/A_0)$

2.23 Welch's Method

- Divide time series into smaller subsets
- May be overlapping
- Apply window (or taper) to each time series
- Calculate power spectrum of subset
- Average all spectra to get smoothed spectrum

(a) 1 DFT taper with NO smoothing

(b) ALL taper with NO smoothing

(c) Welch's method with smoothing and taper

(d) Welch's method with stft and mean

Figure 2.13: Welch's method example

Figure 2.14: Welch Method Combind

2.24 Filtering and Convolution

- Design Filter in Frequency Domain
- Get FT of signal
- Multiply FT of signal with filter
- Inverse FT back to time domain

2.25 Coherency

A standard measure of the similarity of a pair of signals is the coherency function defined by,

$$C(f) = \frac{S_1(f) \cdot S_2(f)}{\sqrt{S_1^2(f) S_2^2(f)}}$$

where S1 and S2 are the complex Fourier transforms of the respective signals and $(\cdot)$ is the dot product. In the case where we have multi-taper estimates of the spectra we can form the coherency function by using all n of the eigenspectra for each signal. In this case the coherence function is calculated by taking the inner vector product of the complex eigenspectra at each frequency,

$$C(f) = \frac{\sum_{k=1}^{n} S_{1k}(f) \cdot S_{2k}^{*}(f)}{\sqrt{S_1^2(f) S_2^2(f)}}$$

where * represents complex conjugation. The coherency function ranges from 0 to 1 and is measure of the coherency at each frequency.

3 Summary

In summary, this book has no content whatsoever.

References

Knuth, Donald E. 1984. "Literate Programming." *Comput. J.* 27 (2): 97–111. https://doi.org/10.1093/comjnl/27.2.97.