Projects in Big Data Analytics

ISYE 4961/6961

Lecture 5

Review: Project 1 – Using Wind Turbine Data

- Import files and extract some data columns into vectors
- Extract column segments
- Build some multi-column arrays
- Display and inspect the data (note: failure occurs at the start of data and time goes backward from there)
- Remove mean and scale, do some correlations and cross plots
- Look at groups of similar sensors, look at pairs of dissimilar sensors
- Look at different time blocks
- Look at same sensor across several turbines
- Report findings

Review: Questions to Answer

- Which sensors are the most important?
- What are the distinguishing features of the data?
- Can we see any indications of failure in the raw data? After mean removal and scaling? In the histograms? In cross plots?
- What does normal operation look like and can we learned anything about regular operational performance?
- Ultimately, can we use the data to anticipate a failure?

Example

Covariance matrix for four wind related sensors

	gen_rpm	rotor_rpm	wind_speed	tower_accel
gen_rpm	1.0000000	0.9999197	0.6800474	0.6364907
rotor_rpm	0.9999197	1.0000000	0.6770360	0.6351792
wind_speed	0.6800474	0.6770360	1.0000000	0.7583575
tower_accel	0.6364907	0.6351792	0.7583575	1.0000000

Are these covariances what you would expect? Why?

Anything seem strange at first glance?

Fourier Transform

```
# transform the scaled and centered values of col 7
>col7_cs_f <- fft(col7_cs)
# calculate the amplitude (modulus or complex spectral magnitude)
>col7_cs_famp <- Mod(col7_cs_f)
# plot the results
>plot(col7_cs_famp[1:50],type=“b”, ,xlab="frequency component",ylab="amplitude")
```

Functions

combine the fft and modulus steps into one

>my_famp <- function(data) {xform <- fft(data); out <- Mod(xform); out}

add in the plot

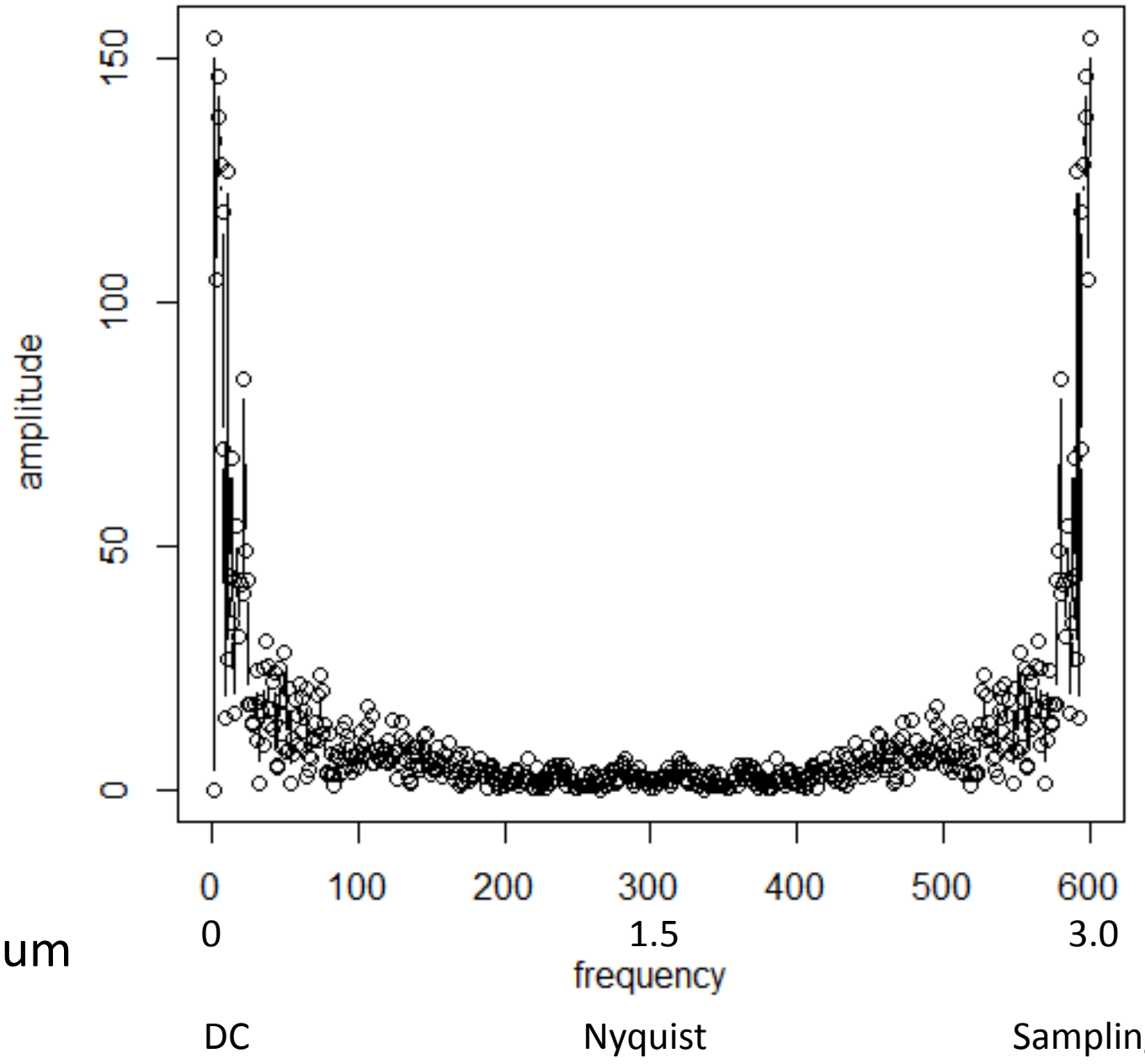
>my_famp <- function(data) {xform <- fft(data); famp <- Mod(xform); plot(famp,type="b",xlab="frequency",ylab="amplitude") }

execute

>my_famp(col7_cs)

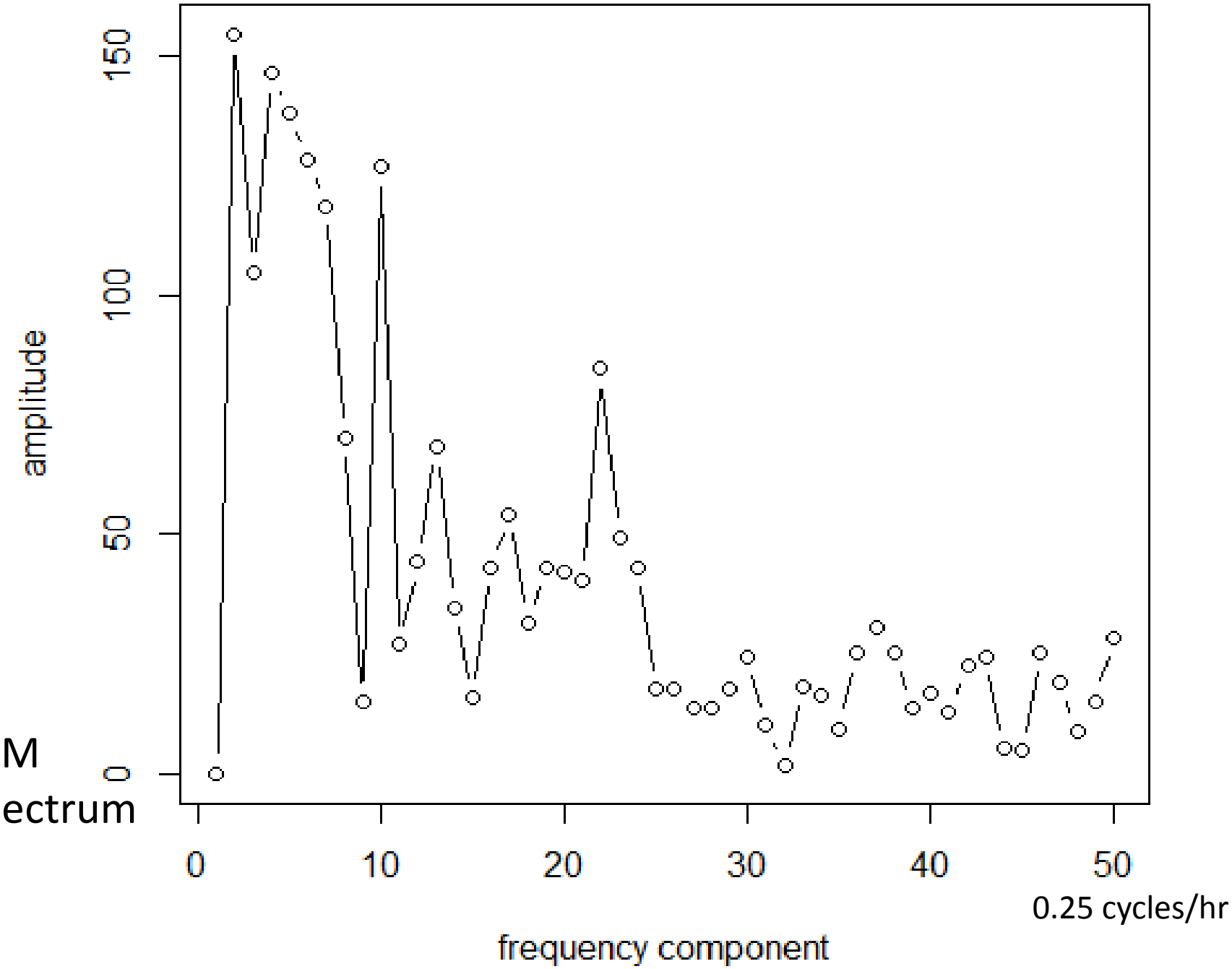
FFT notes

- Scale and remove the mean to the input data. This will produce a zero DC in the Fourier domain and make plotting nicer.
- Use Mod() to get amplitudes and Arg() to get phase.
- Sample spacing in time is 10 min = 600 sec => entire frequency result is 1/600 sec or 0.00167 Hz = .05 cycles/min or 3 cycles/hr. This is the sampling frequency.
- Need to sample at least twice a cycle to avoid aliasing, so highest unaliased frequency is half the sampling frequency or 1.5 cycles /hr.

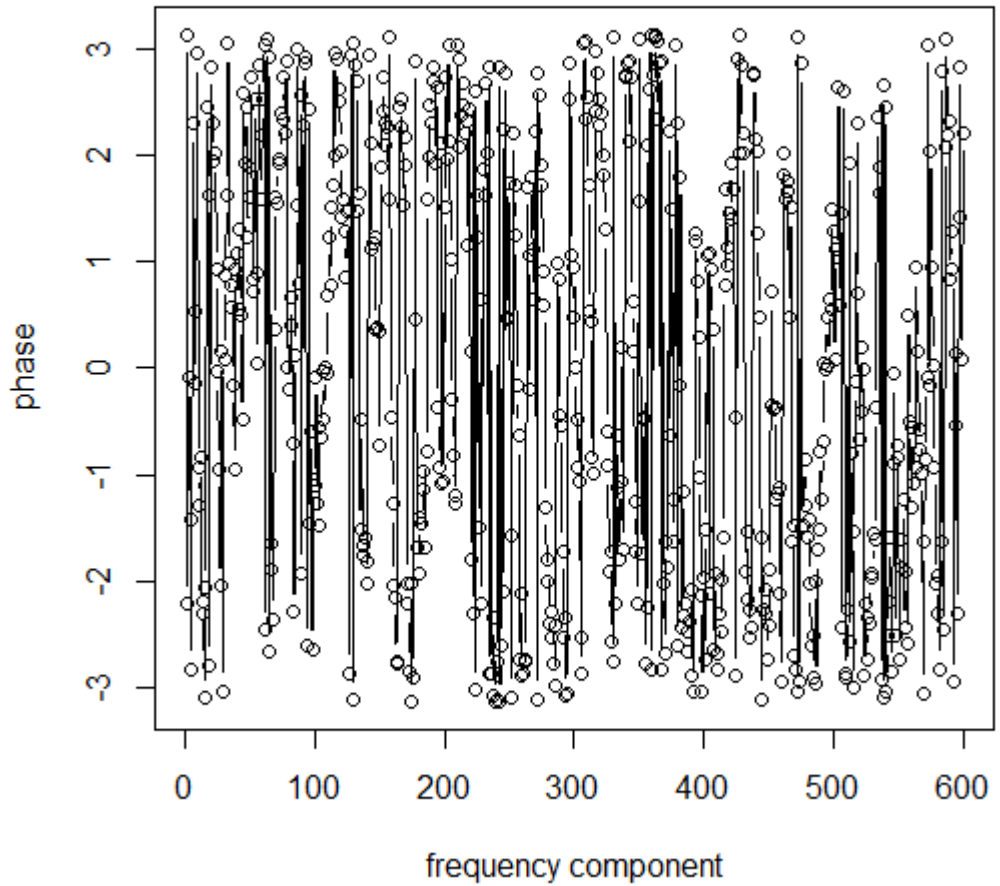


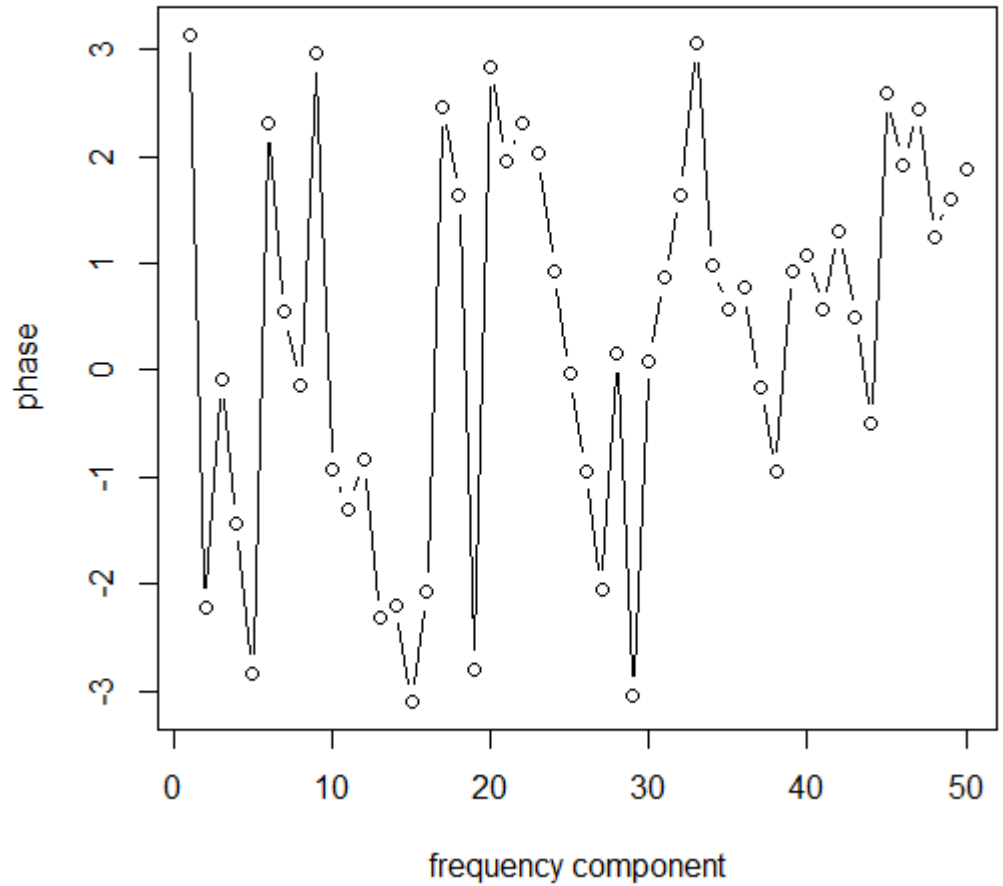
Generator RPM
Frequency Spectrum

Generator RPM
Frequency Spectrum
(detail)



Generator RPM Phase Spectrum





Observations on FFT of generator rpm

- Very slowly varying. All large amplitudes at very low frequencies.
- Even so, there is a characteristic appearance to the amplitude variation.
- With 20 minute sample spacing changes that happen faster than 1.5 cycles per hour will be aliased.
- What can be said about the phase information?