Statistical Approaches to Exoplanetary Science

Eric Feigelson

Center for Astrostatistics

Penn State University

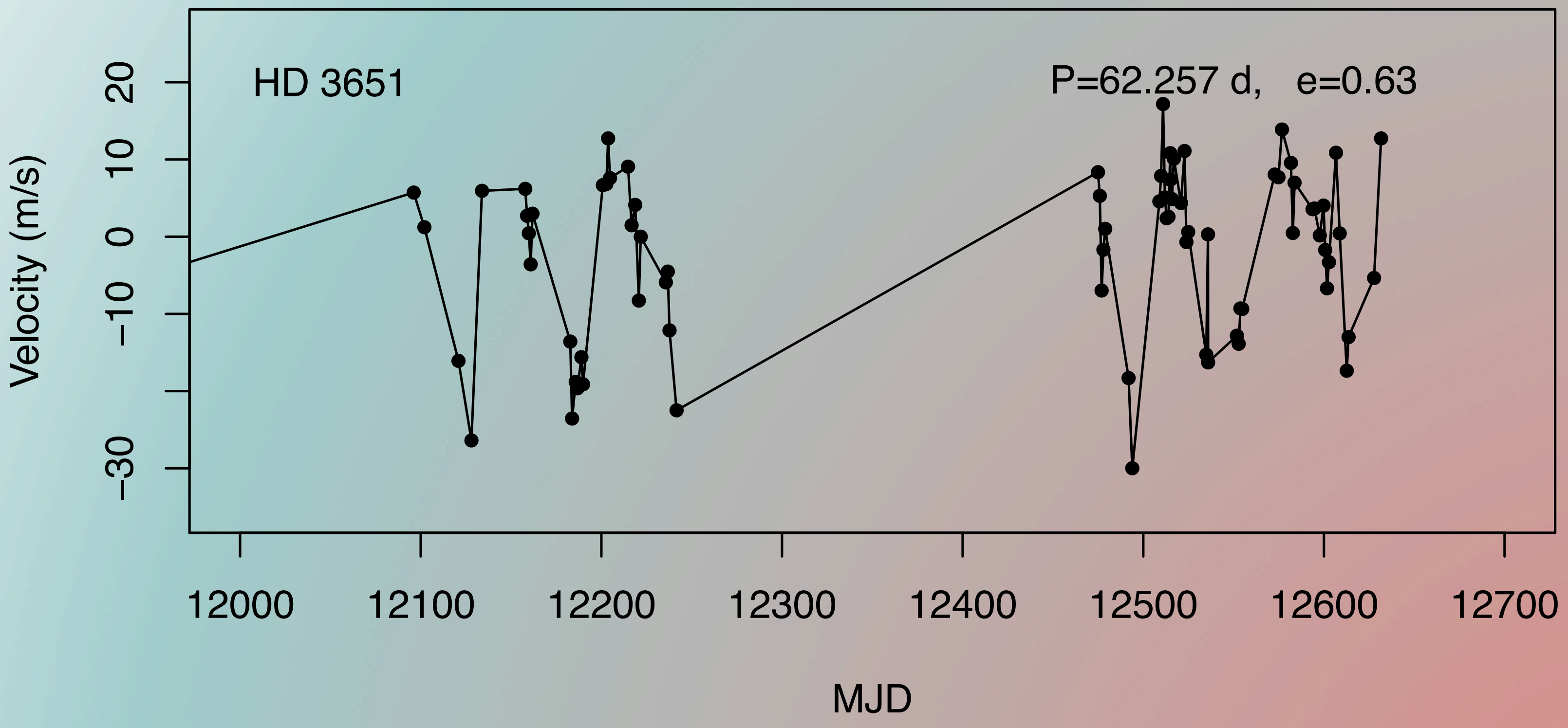
This talk has limited scope:
Time series analysis

for radial velocity and transit studies

- ❖ Nonparametric time domain methods
  - ▪ Nonparametric regression (e.g. Gaussian Processes regression)
- ❖ Parametric time domain methods
  - ▪ ARMA modeling
- ❖ Nonparametric frequency domain methods
  - ▪ Lomb-Scargle periodogram
- ❖ Parametric frequency domain methods
  - ▪ Multiplanet astrophysical modeling

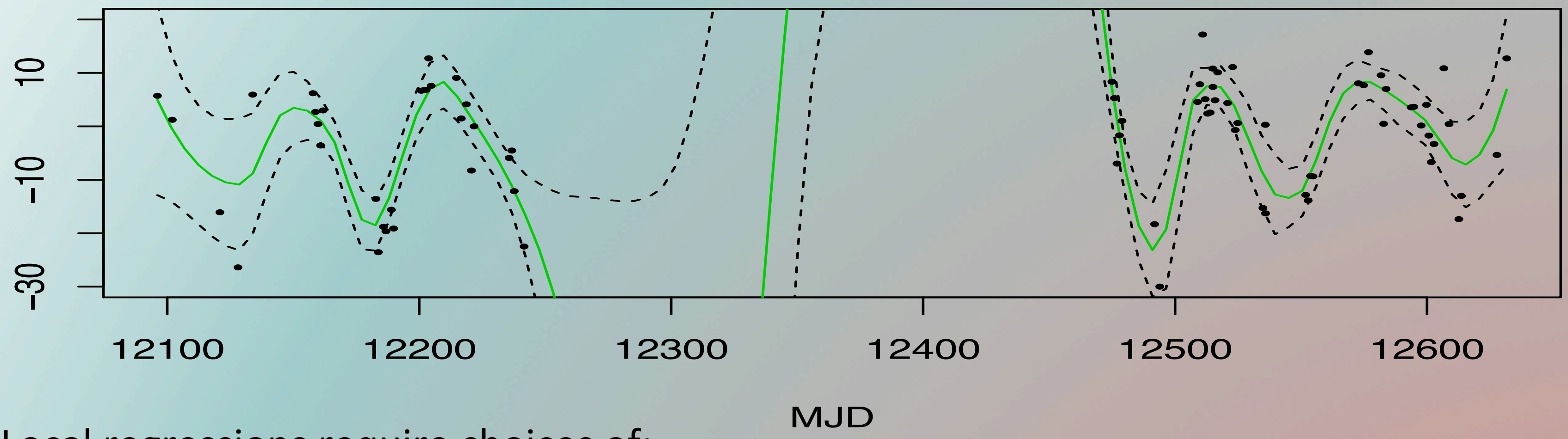
HD 3651

2 years of radial velocity measurements



A Sub-Saturn Mass Planet Orbiting HD 3651
D. Fischer et al. 2003, ApJ 590, 1081

Nonparametric time domain methods

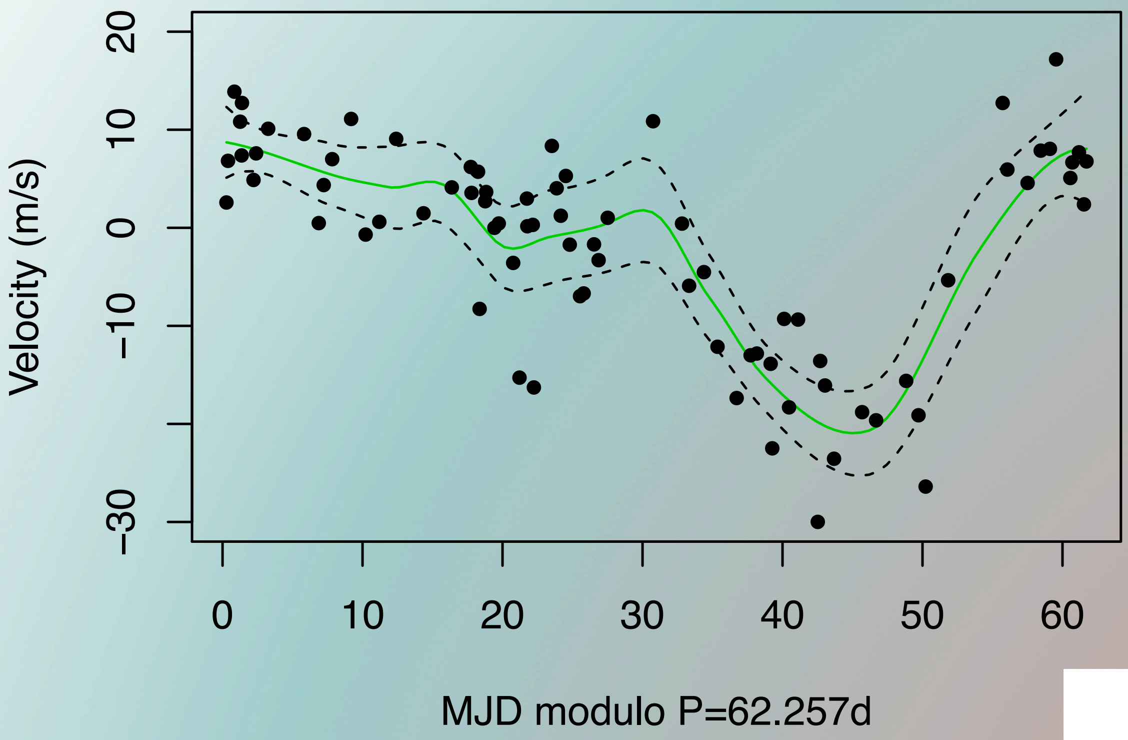


Local regressions require choices of:

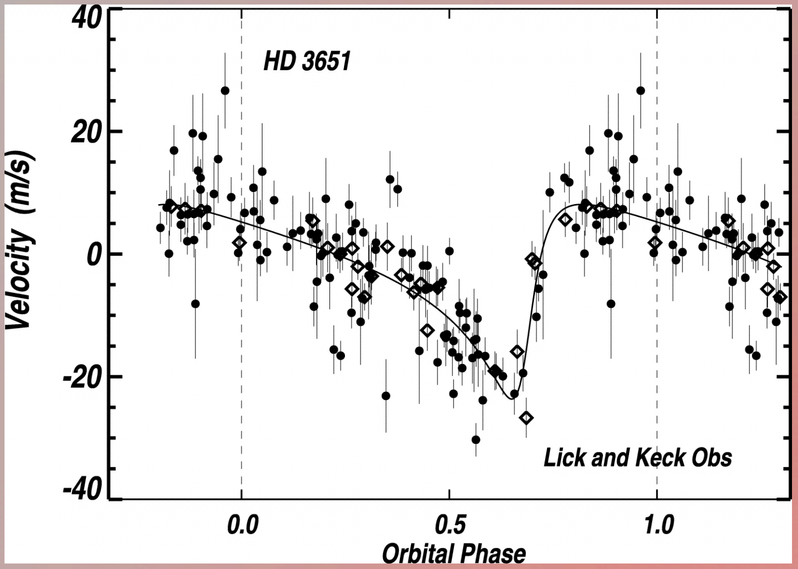
- Weight function: Gaussian, Epanichnikov, …
- Bandwidth: constant, adaptive, AIC d.o.f., …
- Polynomial degree: linear, quadratic, cubic
- Fit criterion: least squares, maximum likelihood
- Confidence bands: local standard deviation, bootstrap

Reference:
Local Regression
C. Loader
Springer 1999

Gaussian processes regression is a local regression estimator that assumes normally distributed residuals. This can be confirmed after the fit is made using a nonparametric test for normality (e.g. Anderson-Darling).



RV modulo orbital period with local fit



More data with parametric astrophysical model (Fischer et al. 2003)

Data and R script for plots

```
   Row    MJD      RV
 [1,] 12096.11   5.72
 [2,] 12102.02   1.23
 [3,] 12120.89 -16.08
 [4,] 12128.08 -26.39
 [5,] 12133.92   5.94
 [6,] 12157.85   6.20
 [7,] 12158.88   2.71
 [8,] 12159.84   0.44
 [9,] 12160.88  -3.58
[10,] 12161.86   2.98
[11,] 12182.81 -13.58
[12,] 12183.80 -23.55
[13,] 12185.80 -18.80
[14,] 12186.80 -19.63
[15,] 12188.95 -15.61
[16,] 12189.83 -19.12
[17,] 12200.82   6.67
[18,] 12201.83   6.77
[19,] 12202.76   6.83
[20,] 12203.76  12.73
[21,] 12204.77   7.59
[22,] 12214.77   9.07
[23,] 12216.73   1.48
[24,] 12218.75   4.12
[25,] 12220.74  -8.27
[26,] 12221.79   0.00
[27,] 12235.70  -5.91
[28,] 12236.76  -4.52
[29,] 12237.73 -12.13
[30,] 12241.64 -22.49
[31,] 12474.92   8.35
[32,] 12475.92   5.29
[33,] 12476.94  -6.97
[34,] 12477.93  -1.69
[35,] 12478.91   1.02
[36,] 12491.87 -18.31
[37,] 12493.92 -29.99
[38,] 12508.91   4.57
[39,] 12509.84   7.87
[40,] 12510.94  17.18
[41,] 12511.96   5.08
[42,] 12512.93   2.39
[43,] 12513.93   2.58
[44,] 12514.92  10.82
[45,] 12515.02   7.38
[46,] 12515.87   4.88
[47,] 12516.92  10.10
[48,] 12520.89   4.36
[49,] 12522.84  11.10
[50,] 12523.88  -0.69
[51,] 12524.86   0.61
[52,] 12534.87 -15.28
[53,] 12535.82   0.30
[54,] 12535.90 -16.28
[55,] 12551.82 -12.82
[56,] 12552.81 -13.87
[57,] 12553.77  -9.29
[58,] 12554.77  -9.36
[59,] 12572.76   8.05
[60,] 12574.83   7.70
[61,] 12576.76  13.88
[62,] 12581.74   9.56
[63,] 12582.79   0.48
[64,] 12583.75   7.00
[65,] 12593.70   3.56
[66,] 12594.75   3.65
[67,] 12597.68   0.16
[68,] 12599.78   4.03
[69,] 12600.72  -1.72
[70,] 12601.71  -6.69
[71,] 12602.78  -3.30
[72,] 12606.67  10.87
[73,] 12608.74   0.43
[74,] 12612.64 -17.36
[75,] 12613.64 -13.00
[76,] 12627.75  -5.35
[77,] 12631.64  12.73
```

Fischer et al. 2003

```
# Radial velocities for HD 3651
# R script, Eric Feigelson, July 2016

rv <- read.table('HD3651_rv.dat')[1:2]
x <- rv[rv[,1]>12000,1]
y <- rv[rv[,1]>12000,2]

plot(x, y, pch=20, type='l', xlab='JD - 2440000', ylab='Velocity (m/s)',
        ylim=c(-30,25))
points(x, y , pch=20)
text(12150, 20, 'HD 3651')
text(12550, 20, 'P=62.257 d,   e=0.63')

install.packages('locfit')  ;  library(locfit)
locfit_model <- locfit(y~lp(x, nn=0.3))
plot(locfit_model, ylim=c(-30,25), band='local', col=3,
        xlab='JD - 2440000', ylab='Velocity (m/s)')
points(xy, pch=20, cex=0.5)

locfit_phase <- locfit(y~lp((x %% 62.257), nn=0.3))
plot(locfit_phase, ylim=c(-30,20), band='local', col=3,
        xlab='JD - 2440000 mod P=62.257d', ylab='Velocity (m/s)')
points(x %% 62.257, y, pch=20)
```

For an R tutorial on nonparametric density estimation see
www2.astro.psu.edu/users/edf/Santiago_2016

Parametric time domain methods

A common task is the search for periodicity from exoplanetary orbits in stellar time series. Nonparametric periodograms include *phase dispersion minimization* (Stellingwerf 1977) and *minimum strength length* (Dworetzky 2003). Parametric periodograms include *box least squares* (Kovacs et al. 2002) for transit detection.

A major impediment is intrinsic stellar variability, usually due to magnetic activity. Local regression can remove trends but is not optimized for stochastic autocorrelated variations. For this, astronomers should be using parametric *autoregressive models* when evenly spaced data is available.

Autoregressive models

Hierarchy of complexity:

AR (autoregressive): current value depends on recent past values

MA (moving average): current change depends on recent past changes

I (integrated): difference operator, removes arbitrary trends

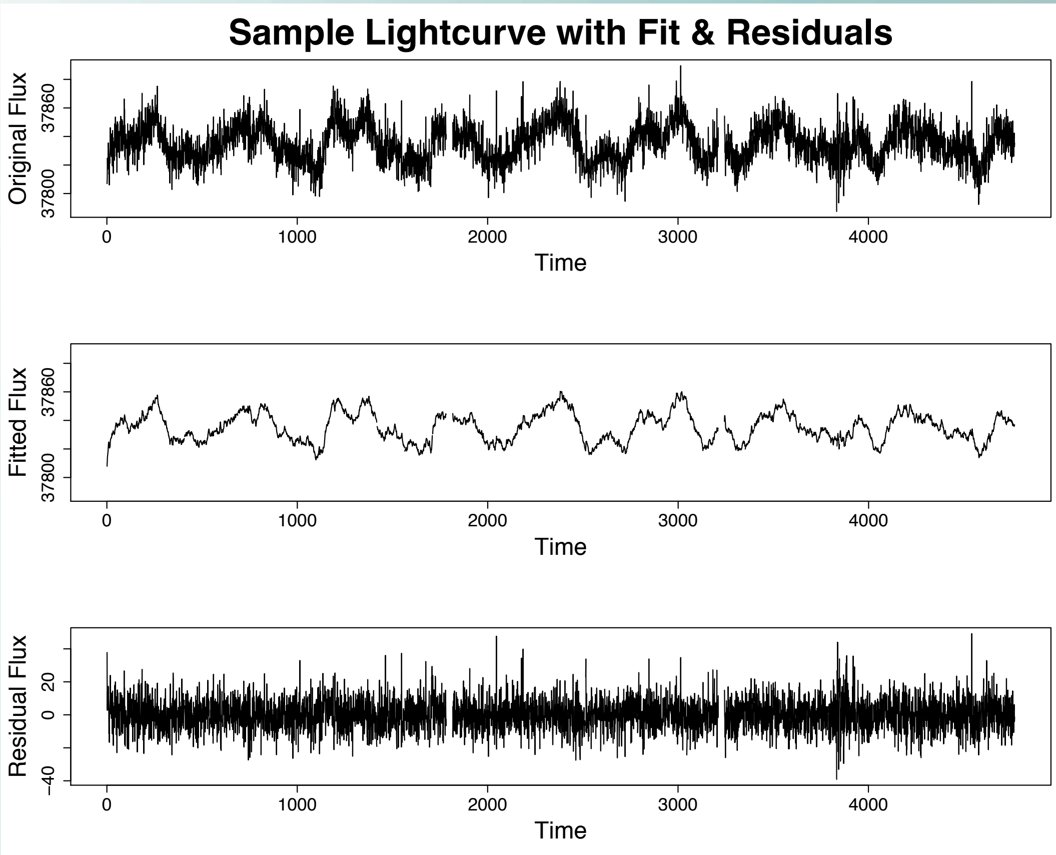
F (fractional): long-memory $1/f^{\alpha}$-type `red noise'

H (heteroscedastic): current variance depends on recent past variances

GARCH: Generalized autoregressive conditional heteroscedastic models, used to predict the volatile stock market

ARFIMA: A powerful family of models treating nonstationarity (trends), short- and long-memory processes. ARFIMA models are extremely effective in reducing correlated variability in Kepler stars (Caceres, Feigelson, et al.)

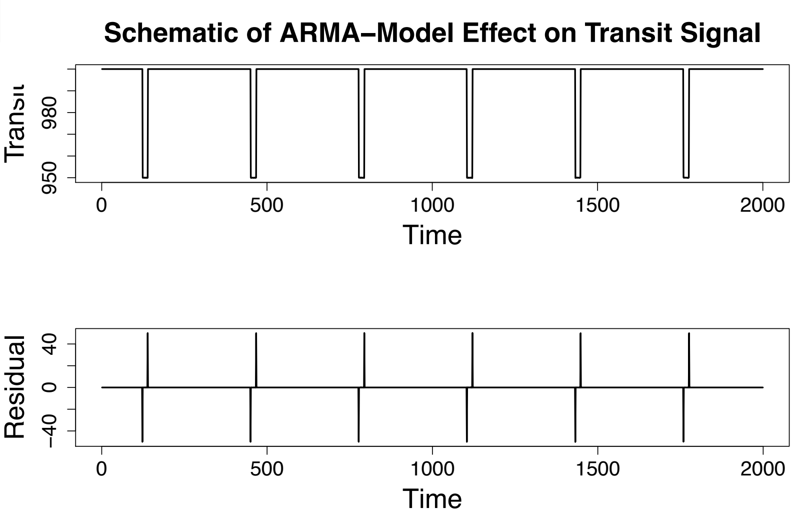
KARPS: Kepler AutoRegressive Planet Search



Typical 4 yr Kepler lightcurve

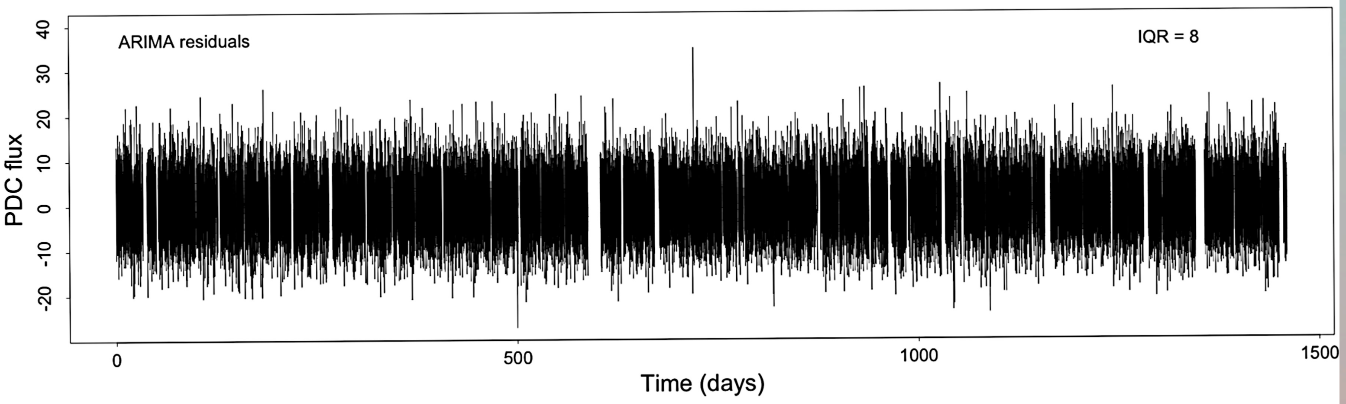
Maximum likelihood ARFIMA model

Residuals

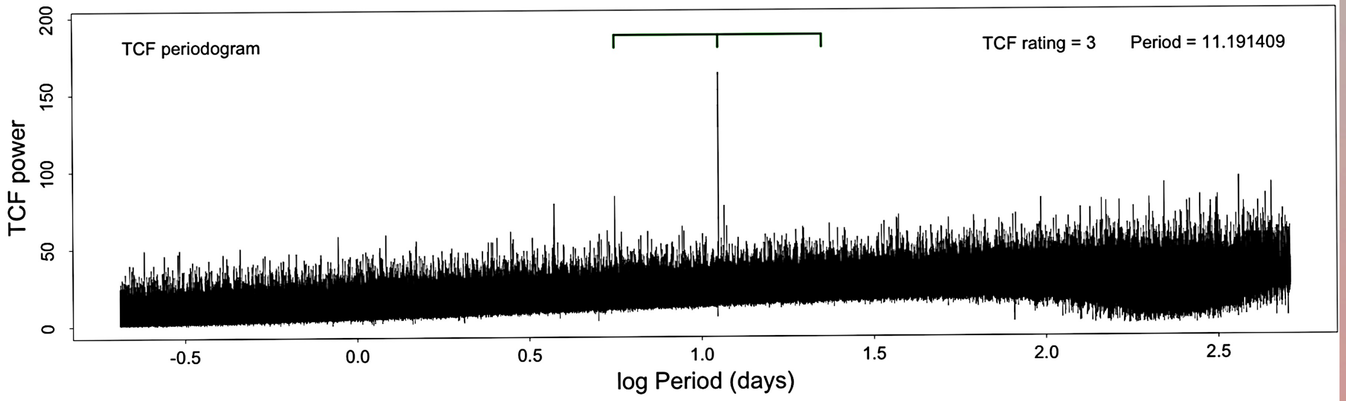


But AR models transform box-shaped transit signal into double-spike signal

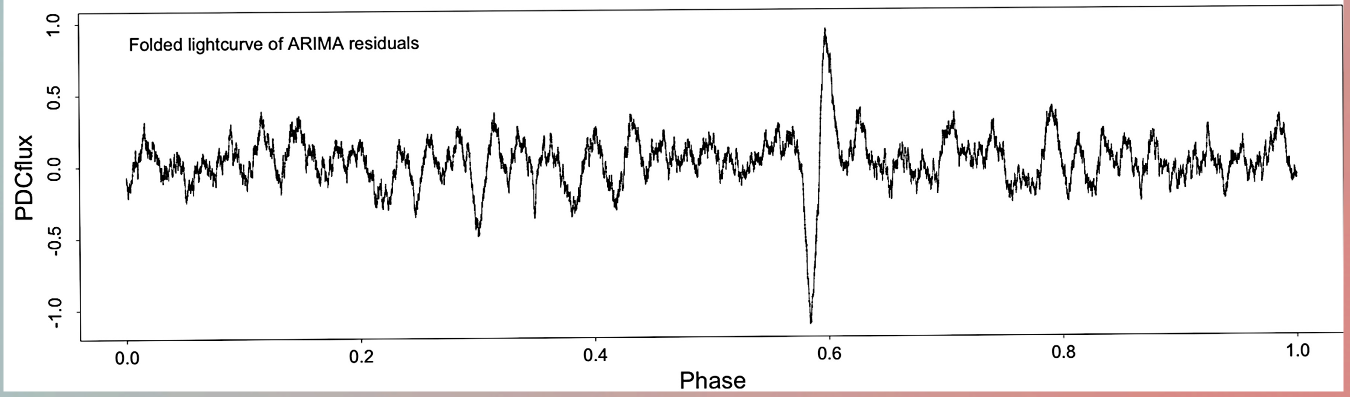
ARFIMA residuals



Transit Comb Filter periodogram

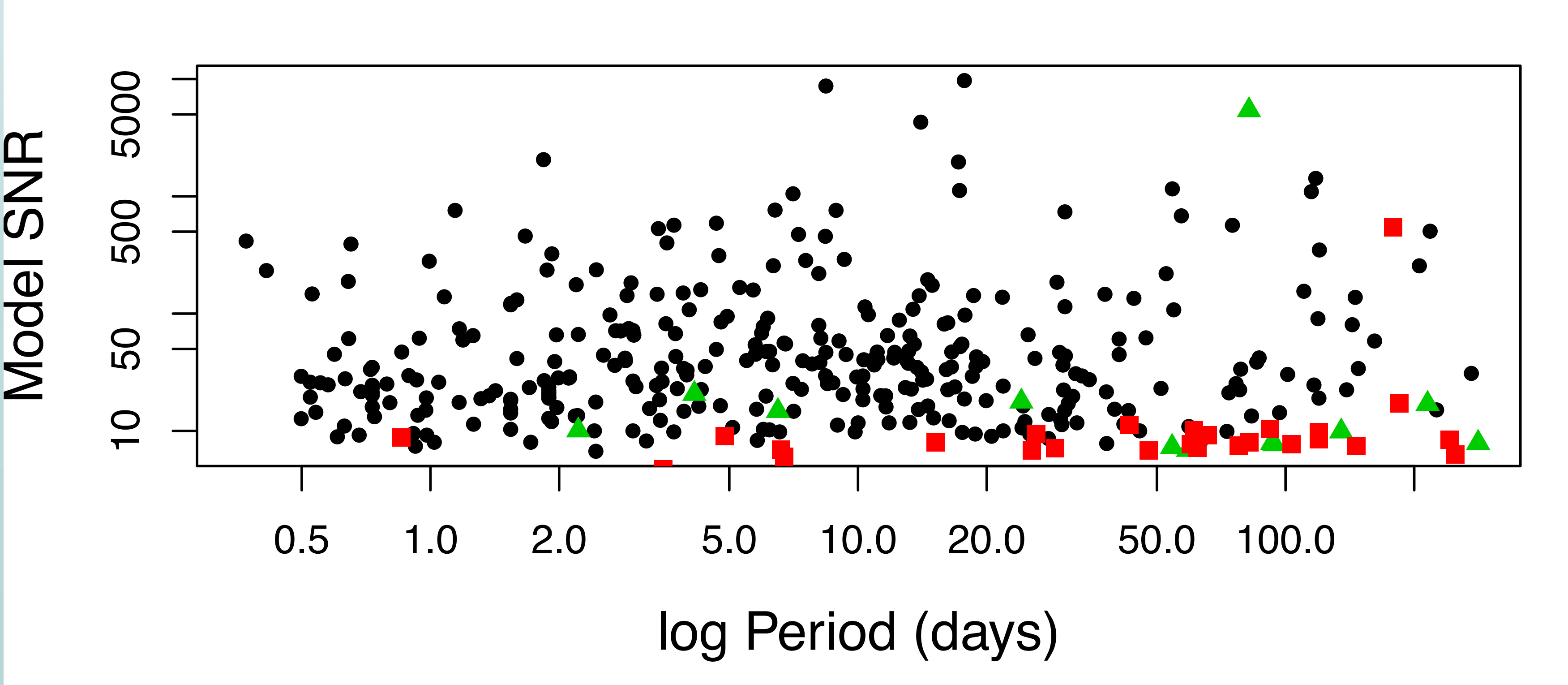


Folded light curve at best TCF period



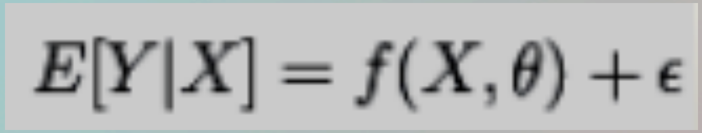
KARPS early results

ARMA+TCF recovers 86% of DR24 Kepler periodic variables (KOIs, black dots). Others are not confirmed (red) or a new period is suggested (green).



Astrophysical time domain methods

Fitting astrophysical models to time domain radial velocity and/or transit time series is very common. These are *nonlinear regression* models:



$$E[Y|X] = f(X, \theta) + \epsilon$$

"The expectation (mean) of the dependent response variable Y for a given value of the indepenent variable(s) X is equal to a specified function *f* which depends both on X and a vector of parameters θ, plus a random error (scatter).

Here, this 2-planet orbital model a 12 parameters in θ: for each planet, the semi-major axis, eccentricity, inclination, ascending node longitude, argument of periastron, and true anomaly

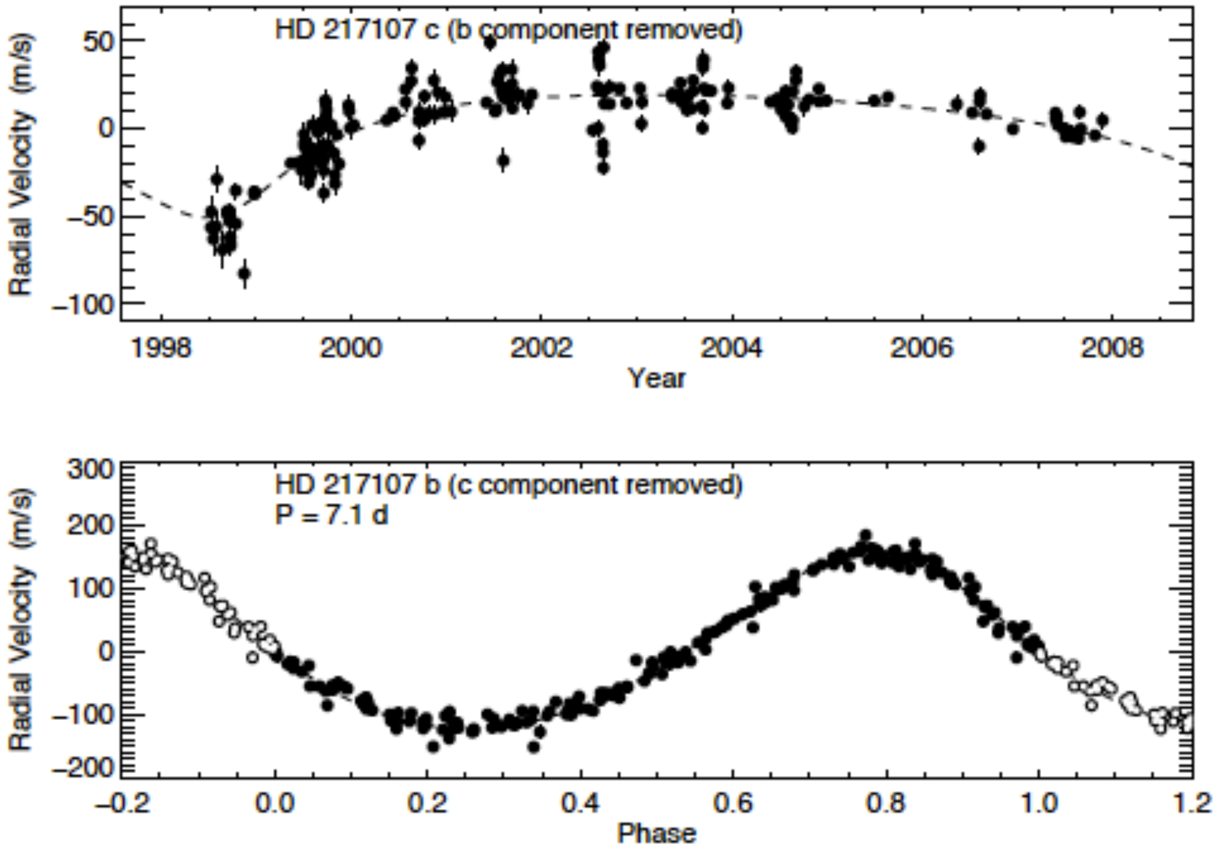
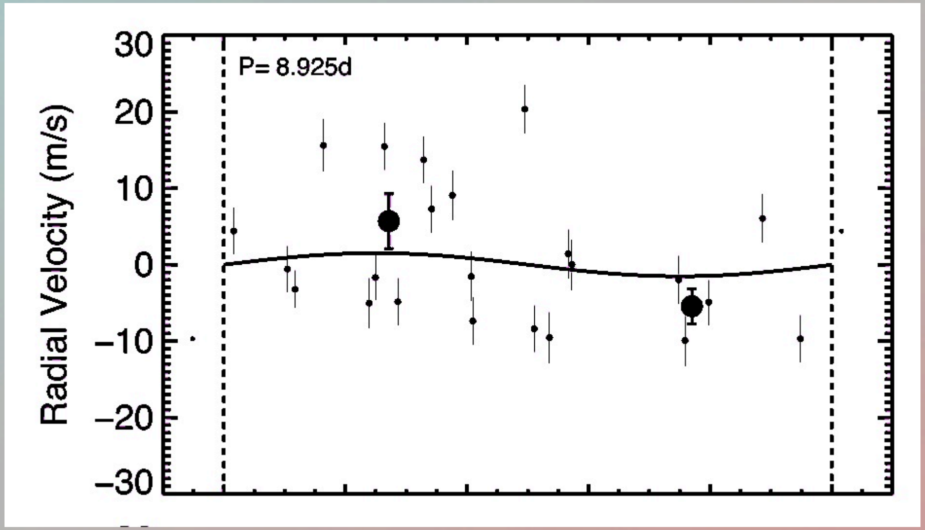


Fig. 7.— RV curves for HD 217107. The data are from Keck Observatory and Lick Observatory, and show the inner planet (top) with $P = 7.1$ d and $M \sin i = 1.4\mathrm{M}_{\mathrm{Jup}}$ and the outer planet (bottom) with $P = 11.6$ yr and $M \sin i = 2.6\ \mathrm{M}_{\mathrm{Jup}}$.

Wright et al. (2009)

Problems with regression in the astronomical literature

- **Improper use of minimum χ^2 fitting weighted by measurement errors** Valid only if measurement errors are the sole source of scatter. Difficulty with model selection.



- **Inadequate residual analysis** Variance fraction? (adjusted R^2 & Mallow's C_p) Goodness-of-fit (Anderson-Darling test) Structure? (Local regression) Autocorrelated? (Durbin-Watson test, ARMA model) Normally distributed? (Anderson-Darling test, quantile regression) Outliers? (Cook's distance plot)

Inadequate model validation Goodness-of-fit test with Kolmogorov-Smirnov or (more sensitive) Anderson-Darling nonparametric 1-sample tests

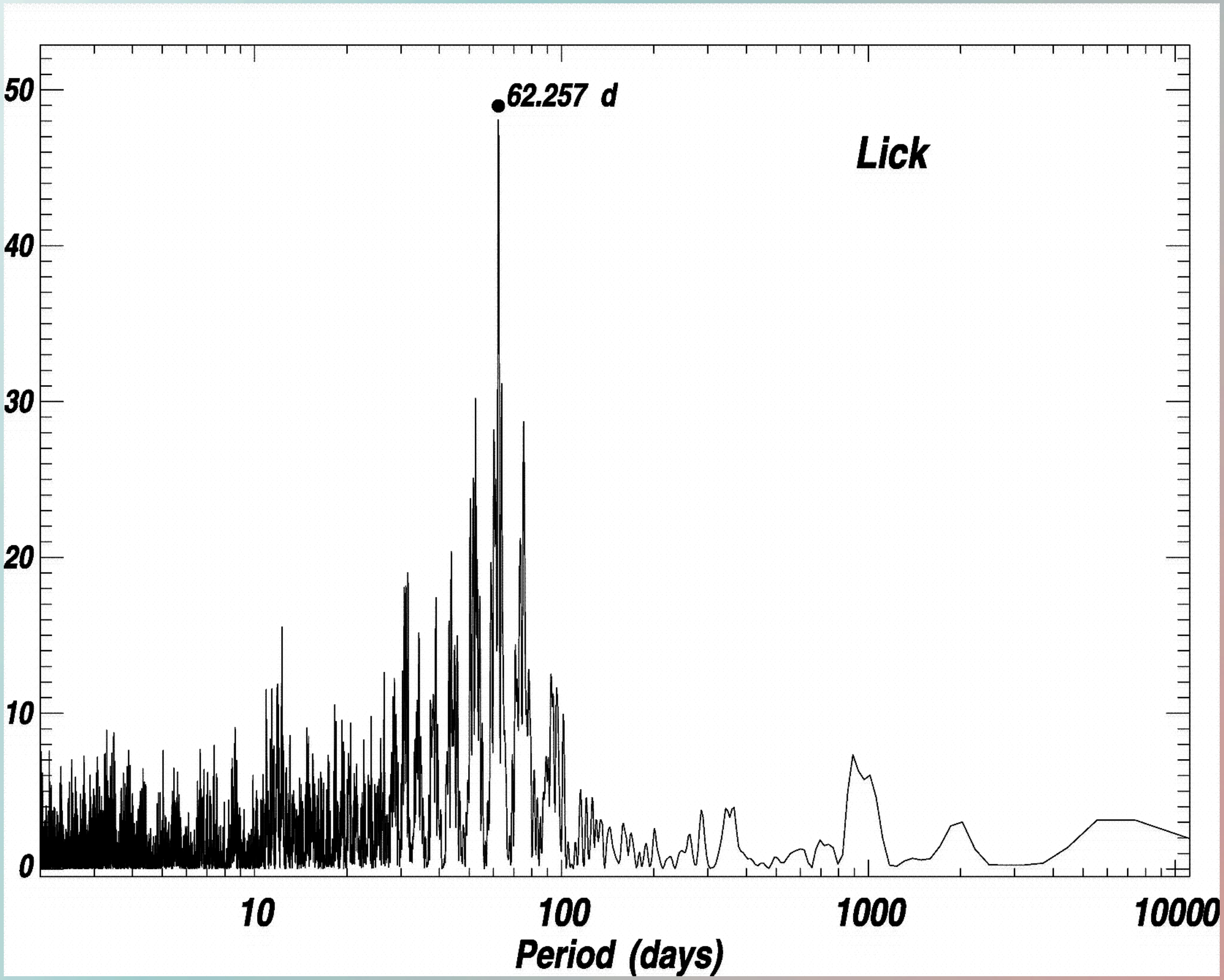
Weak model selection Penalized likelihood measures for model elaboration and parsimony

Underuse of multivariate regression Not a sequence of bivariate analyses

Sensitivity to arbitrary transformations Science result should not depend on units or log-transform

Overuse of Bayesian inference If no prior information is available and science is based on the mode (maximum) of the posterior, then it is maximum likelihood estimation. Either MCMC or standard optimization methods (steepest descent, simplex, EM, …) can be used.

Nonparametric frequency domain methods



A Fourier periodogram for HD 3651

Fourier analysis

Spectral analysis reveals nothing of the evolution in time, but rather reveals the variance of the signal at different frequencies. The *power spectral density* or *periodogram* is the modulus squared of the Fourier transform of a time series (Fourier 1807, Schuster 1898).

Fourier analysis is valid only under restrictive assumptions: an infinitely long dataset of equally-spaced observations of a stationary process consisting of homoscedastic Gaussian noise with purely periodic signals of sinusoidal shape. Even the it is formally an *inconsistent* estimator.

The Lomb-Scargle periodogram is a generalized Fourier periodogram for unevenly spaced data widely used in astronomy.

Recommended text:
D. B. Percival and A. T. Walden (1992)
Spectral Analysis for Physical Applications

Improving the periodogram

The signal-to-noise of a periodogram can be improved by *smoothing* (in the frequency domain). *Tapering* (in the time domain) that shrinks signal amplitude at the ends of the time series reduces spectral leakage.

Filtering the time domain data prior to spectral analysis is also helpful, particular removal of aperiodic long-term trends and autoregressive behaviors using local regression and ARMA techniques.

Astronomers can benefit from filtering their time series, smoothing and (multi)tapering their Fourier/L-S periodograms

However, these steps do not alleviate aliasing of periodicities due to unevenly spaced observations

False Alarm Probabilities

It is extremely difficult to derive the significance of a weak periodicity from any type harmonic analysis. This is particularly true for unevenly spaced data with a nonrandom cadence. ***Do not trust analytical estimates (P ~ exp(d_ν/σ^2) for FAPs***, as their mathematical derivation is very restrictive and rarely applies to real data (e.g. Koen, MNRAS 1990).

I believe it is essential to make simulations keeping the observing times fixed:

- Permute or bootstrap the data after trend removal
- If autocorrelation is present (Durbin-Watson test), the process should be characterized(ARMA modeling) and simulate.
- If a periodicity may be present, it should be simulated.

The ensemble of simulated periodograms (Fourier, L-S, PDM, MSL, …) can then be compared to the observed periodogram.

Final comments

Astronomers are often familiar with only a narrow suite of time series procedures. A vast methodology has been developed for signal processing & econometrics. Unfortunately, little applies to irregularly spaced observations.

Exoplanetary research has an acute need for powerful time series and regression methods. Knowledge and use of sophisticated statistical concepts and methods can improve the reliability of our scientific results.