A Search for Habitable Planets

Data reduction strategies and challenges: Analysis approach and Transit detection

Jon M. Jenkins
Co-I Data Analysis

SETI Institute
NASA Ames Research Center

Michelson Summer Workshop

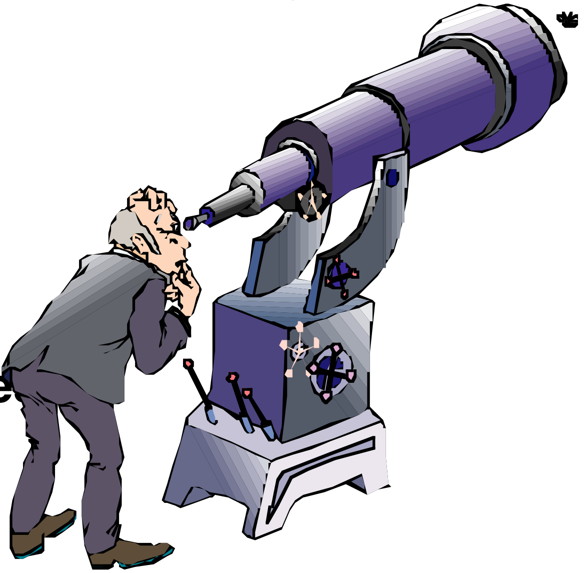
24 July 2007

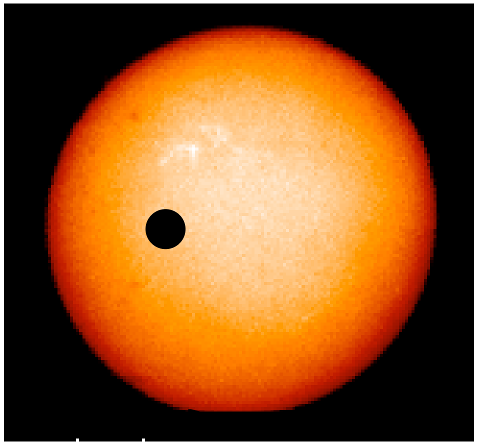
SAO

STScI

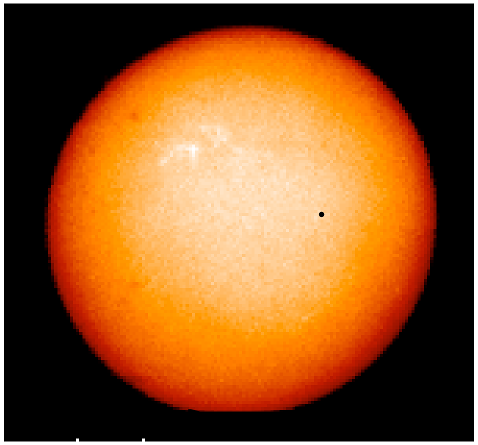
OVERVIEW

- How Do We Best Detect Transiting Planets?
- Simple Matched Filters and Their Properties
- Colored Noise and Resulting Matched Filters
- A Wavelet-Based Approach
- Establishing Confidence in Transit Candidate
- Detecting Multiple Transiting Planets
- Centroiding for False Positive Elimination





Jupiter:
1% area of the Sun (1/100)



Earth or Venus
0.01% area of the Sun (1/10,000)

The Challenge for Kepler

Detect Earth-sized Transits (80 ppm or 0.11 W/m^2 reduction in brightness) against:

Stellar Variability + Shot Noise + Instrument Noise

A search through 100,000 Target Stars & Orbital Periods up to 2 years requires

$\sim 1 \times 10^{12}$ statistical tests

$\Rightarrow$ total SNRs of $\geq 7\sigma$ are required

$\therefore$ Single Event SNR's of $\geq 3.5\sigma$ are required for 4 transits

Shot Noise = 1.4×10^{-5} in 6.5 hours

Instrument Noise* = 0.7×10^{-5} in 6.5 hours

Detecting Deterministic Signals

The problem:

- H0: $x(n)= w(n)$ or
- H1: $x(n)= s(n) + w(n)$

$s(n)$ is the signal of interest

$x(n)$ is the time series we observe

$w(n)$ is the observation noise (Gaussian)

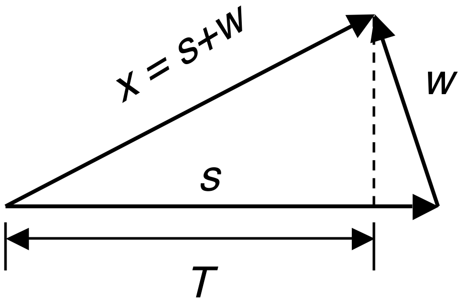
The best method for detecting a known signal in additive Gaussian noise is a matched filter

A matched filter measures the correlation between the data and the signal, normalized by the rms variation of the observation noise

(Simple) Matched Filters

Define

$$T = \frac{x^T s}{\sigma_w \sqrt{s^T s}}$$



Under H0:

$$\langle T \rangle = 0, \quad \sigma_T^2 = 1$$

Under H1:

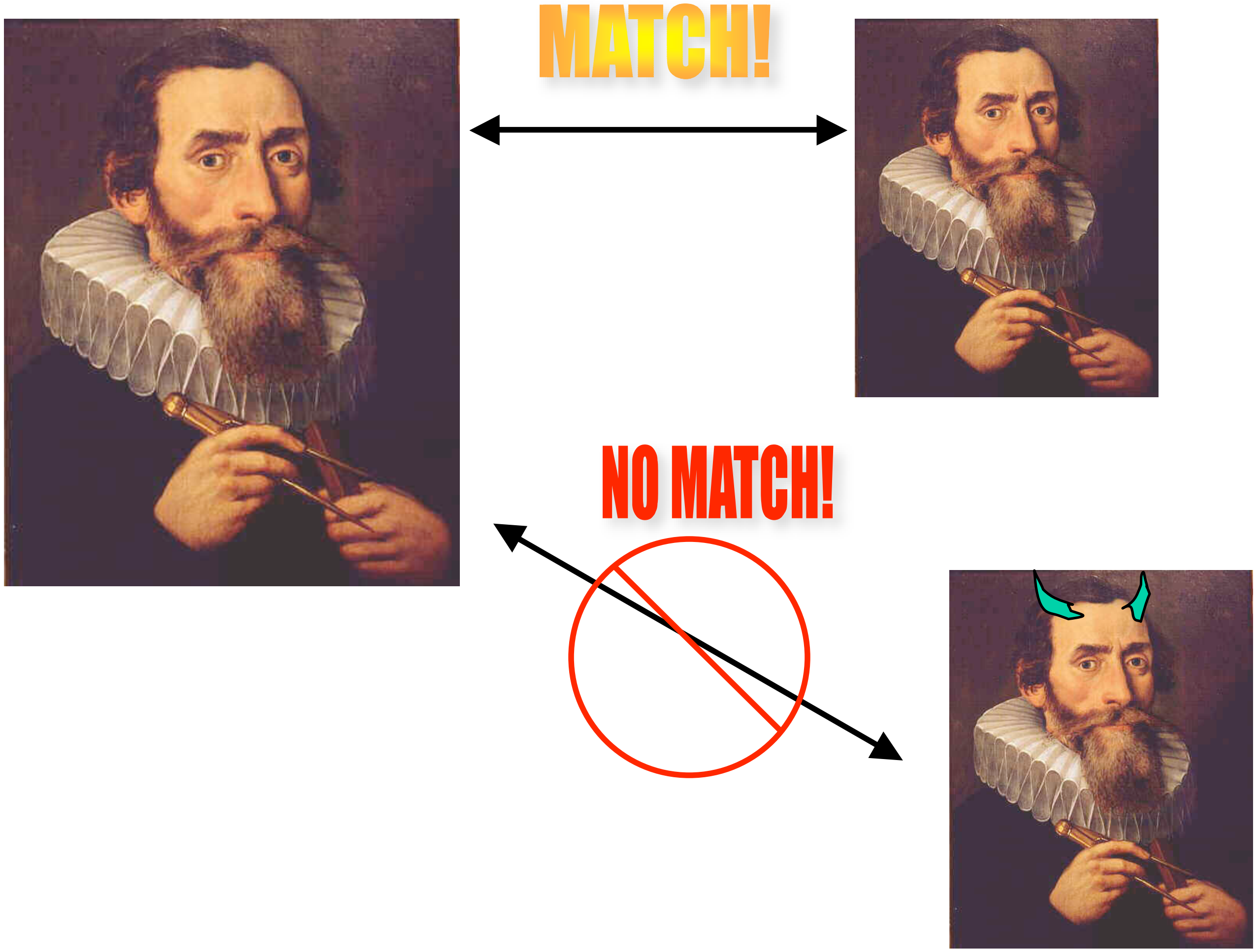
$$\langle T \rangle = \tfrac{1}{\sigma_w} \sqrt{s^T s}, \quad \sigma_T^2 = 1$$

If $T < \gamma$, then choose H0, if $T > \gamma$, then choose H1

Note: Least-Squares Amplitude=$s^T x/s^T s$

Change in χ^2: $\Delta\chi^2 = T^2$

What Does This Mean?



MATCH!

NO MATCH!

Properties of Detection Statistic *T*

Under H0:

$$\langle T\rangle_{\mathbf{H0}} = \left\langle \frac{\mathbf{w}^{\mathrm{T}}\mathbf{s}}{\sigma_w\sqrt{\mathbf{s}^{\mathrm{T}}\mathbf{s}}} \right\rangle = \frac{\langle \mathbf{w}^{\mathrm{T}}\rangle \mathbf{s}}{\sigma_w\sqrt{\mathbf{s}^{\mathrm{T}}\mathbf{s}}} = \frac{0^{\mathrm{T}}\mathbf{s}}{\sigma_w\sqrt{\mathbf{s}^{\mathrm{T}}\mathbf{s}}} = 0$$

$$\left\langle \left(T - \langle T\rangle\right)^2 \right\rangle_{\mathbf{H}0} = \left\langle \left[\frac{\mathbf{w}^{\mathrm{T}}\mathbf{s}}{\sigma_w\sqrt{\mathbf{s}^{\mathrm{T}}\mathbf{s}}} \right]^2 \right\rangle = \frac{\mathbf{s}^{\mathrm{T}}\langle \mathbf{w}^{\mathrm{T}}\mathbf{w}\rangle \mathbf{s}}{\sigma_w^2\,\mathbf{s}^{\mathrm{T}}\mathbf{s}} = \frac{\sigma_w^2\mathbf{s}^{\mathrm{T}}\mathbf{s}}{\sigma_w^2\,\mathbf{s}^{\mathrm{T}}\mathbf{s}} = 1$$

–

Under H1:

$$\langle T\rangle_{\mathbf{H1}} = \left\langle \frac{(\mathbf{w}+\mathbf{s})^{\mathrm{T}}\mathbf{s}}{\sigma_w\sqrt{\mathbf{s}^{\mathrm{T}}\mathbf{s}}} \right\rangle = \frac{\left(\langle \mathbf{w}\rangle + \mathbf{s}\right)^{\mathrm{T}}\mathbf{s}}{\sigma_w\sqrt{\mathbf{s}^{\mathrm{T}}\mathbf{s}}} = \frac{\mathbf{s}^{\mathrm{T}}\mathbf{s}}{\sigma_w\sqrt{\mathbf{s}^{\mathrm{T}}\mathbf{s}}} = \frac{\sqrt{\mathbf{s}^{\mathrm{T}}\mathbf{s}}}{\sigma_w}$$

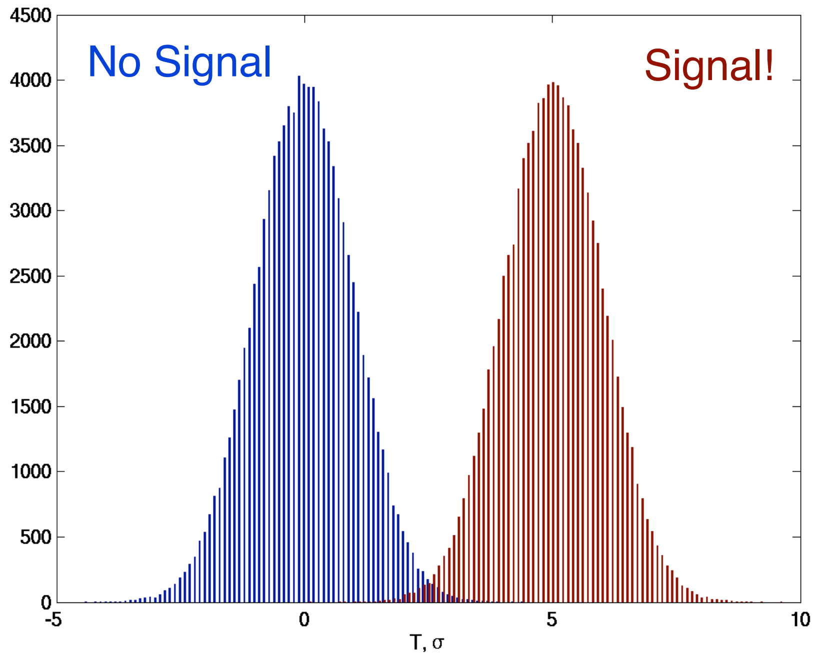
$$\left\langle \left(T - \langle T\rangle\right)^2 \right\rangle_{\mathbf{H1}} = \left\langle \left[\frac{(\mathbf{w}+\mathbf{s})^{\mathrm{T}}\mathbf{s}}{\sigma_w\sqrt{\mathbf{s}^{\mathrm{T}}\mathbf{s}}} - \frac{\sqrt{\mathbf{s}^{\mathrm{T}}\mathbf{s}}}{\sigma_w} \right]^2 \right\rangle = \frac{\left\langle \left[\mathbf{w}^{\mathrm{T}}\mathbf{s} + \mathbf{s}^{\mathrm{T}}\mathbf{s} - \mathbf{s}^{\mathrm{T}}\mathbf{s}\right]^2 \right\rangle}{\sigma_w^2\mathbf{s}^{\mathrm{T}}\mathbf{s}} = \frac{\sigma_w^2\mathbf{s}^{\mathrm{T}}\mathbf{s}}{\sigma_w^2\mathbf{s}^{\mathrm{T}}\mathbf{s}} = 1$$

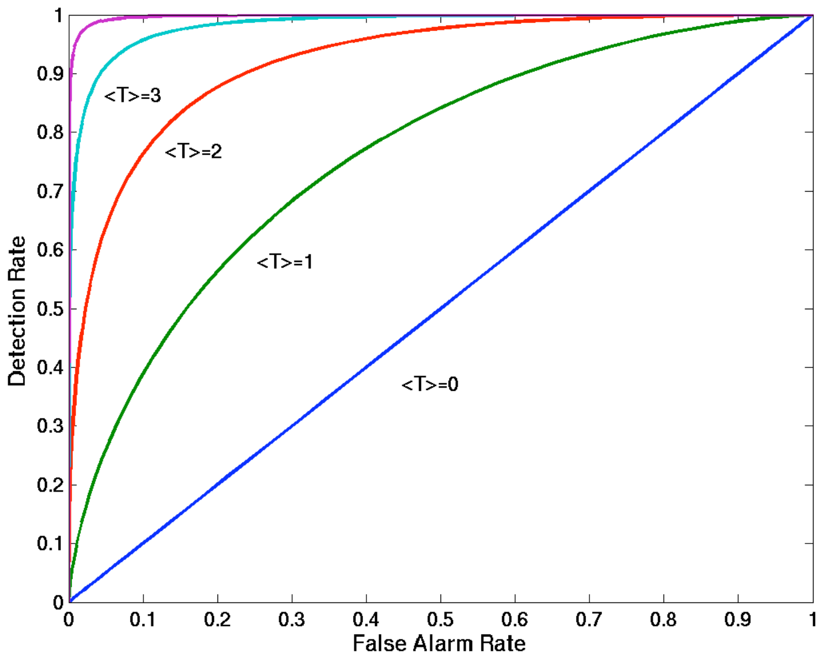
Receiver Operating Curves

T is a Gaussian random variable

$$P_F = \frac{1}{\sqrt{2\pi}} \int_{\gamma}^{\infty} \exp\left(-\tfrac{1}{2} y^2\right) dy$$

$$P_D = \frac{1}{\sqrt{2\pi}} \int_{\gamma-\langle T\rangle}^{\infty} \exp\left(-\tfrac{1}{2} y^2\right) dy$$





How do we choose the threshold, γ?

If amplitude of *s* not known, we set γ to control $P_{F.}$
(Neyman-Pearson Criterion)

Colored Noise

Let w(n) be Gaussian, but not necessarily white, with autocorrelation matrix *R*

Now, decide H1 if

$$T = \frac{x^T R^{-1} s}{\sqrt{s^T R^{-1} s}} > \gamma$$

Now, using SVD, $R = U \cdot S \cdot V^T$**, so that** $R^{-1} = V^T \cdot S^{-1} \cdot U$

Moreover, $U = V$ **so that** $R^{-1} = U^T \cdot S^{-1/2} \cdot S^{-1/2} \cdot U = H^T \cdot H$

Thus,

$$T = \frac{(Hx)^T (Hs)}{\sqrt{(Hs)^T (Hs)}} = \frac{\tilde{x}^T \tilde{s}}{\sqrt{\tilde{s}^T \tilde{s}}} > \gamma$$

So optimal detector is a prewhitener followed by a simple matched filter

Colored Noise

How do we determine R?

If the noise is stationary, we can work in the frequency domain:

Decide H1 if

$$T = {T'}\big/{\overline{T}} = \int \frac{X(f)S^{*}(f)}{P(f)}df \Bigg/ \sqrt{\int \frac{S(f)S^{*}(f)}{P(f)}df} > \gamma$$

Note:

$$\overline{T} = \sqrt{\int \frac{S(f)S^{*}(f)}{P(f)}df}$$

$\overline{T}$ **is the average detection statistic.**

And CDPP = Transit Depth ÷ $\overline{T}$

PSDs for Solar-Like Variability

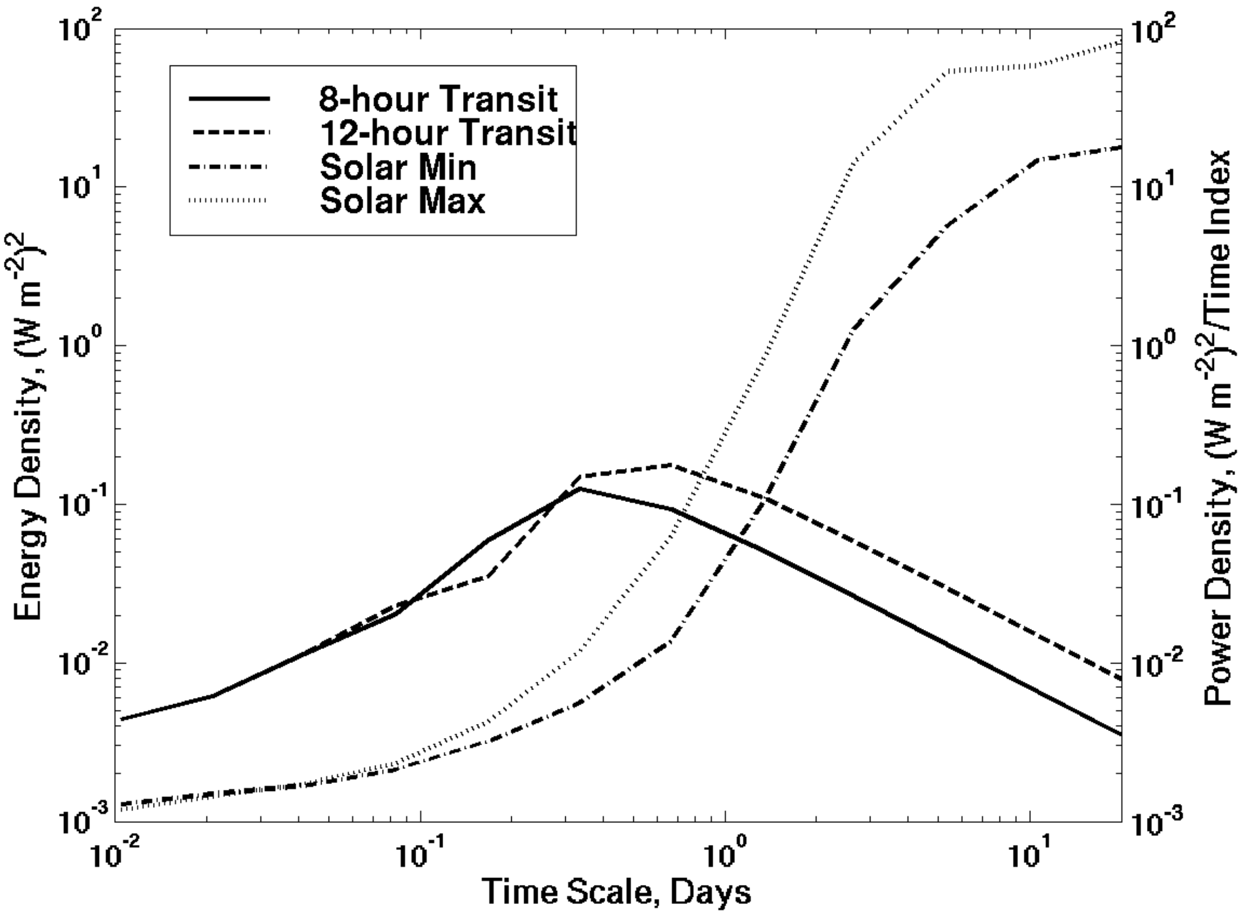
Is stellar variability stationary?

No!

Alternative solution:

Work in a joint time-frequency domain

Wavelets are a natural choice



Why Wavelets?

Several Choices to estimate *R* or *P(ω):*

Time Domain Techniques (eg., Model *w(n)* as an ARMA process)

- Good resolution in time, data gaps can be accommodated
- Resolution in frequency cannot in general be controlled

Periodogram-Based Techniques (eg., STFT, Kay 1999)

- Good resolution in frequency, poor resolution in time
- Analysis of all timescales occurs with same length window
- Smooth adaptation is numerically intensive

Wavelets (Jenkins 2002)

- Resolution in time can be balanced with resolution in frequency
- Different timescales can be analyzed independently
- Numerically efficient for joint time-frequency analysis
- Dyadic wavelet passbands well-matched to *1/f*-type processes

A Wavelet-Based Approach

The time series *x(n)* is partitioned (filtered) into complementary channels

$$W_X(i,n) = \{h_1(n) * x(n), h_2(n) * x(n),\ldots, h_M(n) * x(n)\}$$
$$= \{x_1(n), x_2(n),\ldots, x_m(n)\}$$

Properties of OWT's:

1. Linear

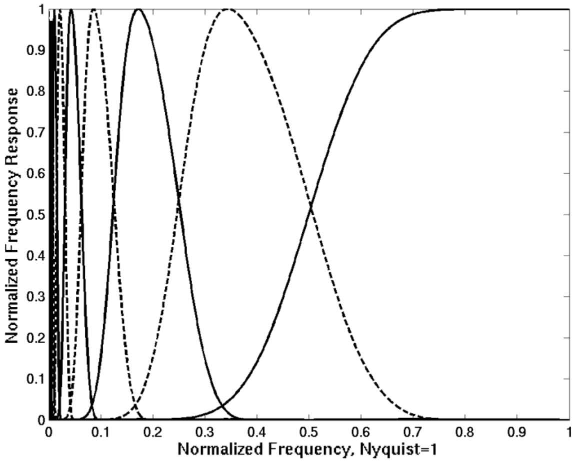
$$W\{a\,x(n)+b\,y(n)\} = a\,W\{x(n)\} + b\,W\{y(n)\}$$

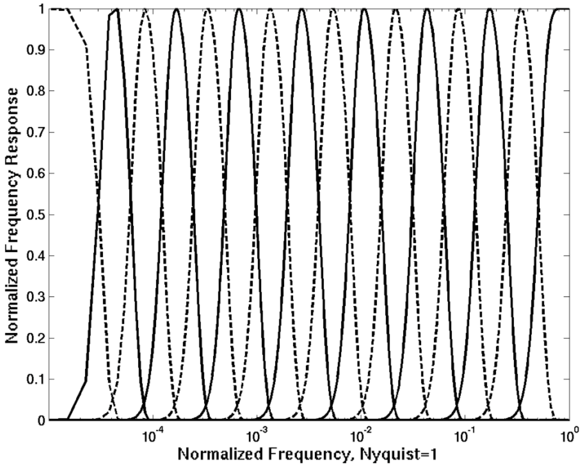
2. Shift-Invariant (does *not hold* for critically sampled wavelets)

$$W(x(n-k)) = \{\, x_1(n-k), x_2(n-k), \ldots, x_M(n)\,\}$$

3. Scalar Products (Parseval's Relation):

$$x\cdot y = 2^{-1}x_1\cdot y_1 + 2^{-2}x_2\cdot y_2 + \ldots + 2^{-(M-1)}x_{M-1}\cdot y_{M-1} + 2^{-(M-1)}x_M\cdot y_M$$





A Wavelet-Based Matched Filter (Text Version)

Implement Adaptive Matched Filter in Wavelet Domain:

1. Transform data into wavelet domain W_x

2. Transform transit pulse into wavelet domain W_s

3. Estimate noise power in each channel $\{\sigma^2_1(n), \sigma^2_2(n), \ldots, \sigma^2_M(n)\}$

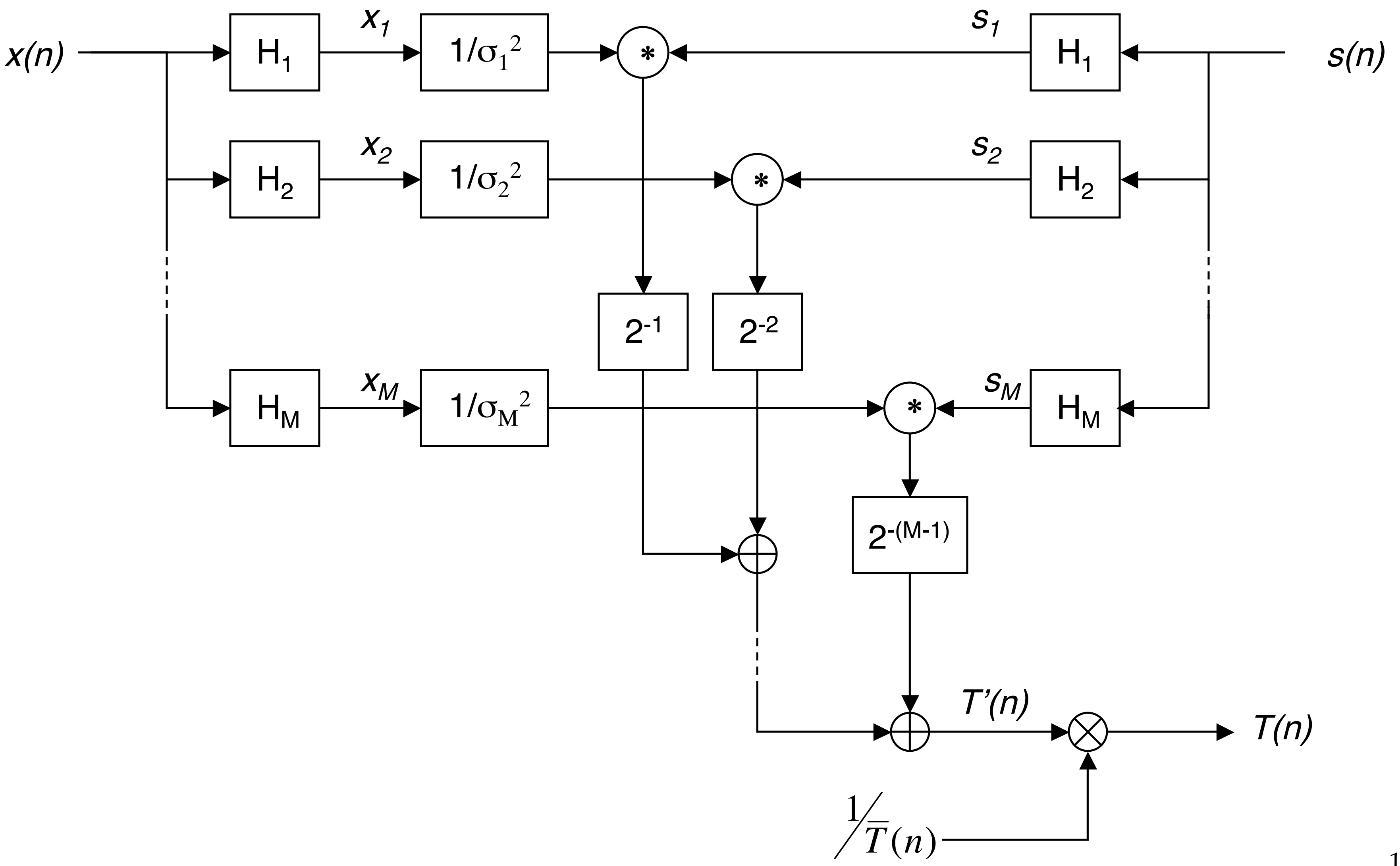
4. Divide each $x_i(n)$ pointwise by $\sigma^2_\iota(n)$

5. Convolve doubly-whitened data with transit pulse wavelets $[x_i(n)/\sigma^2_\iota(n)] * s_i(n)$

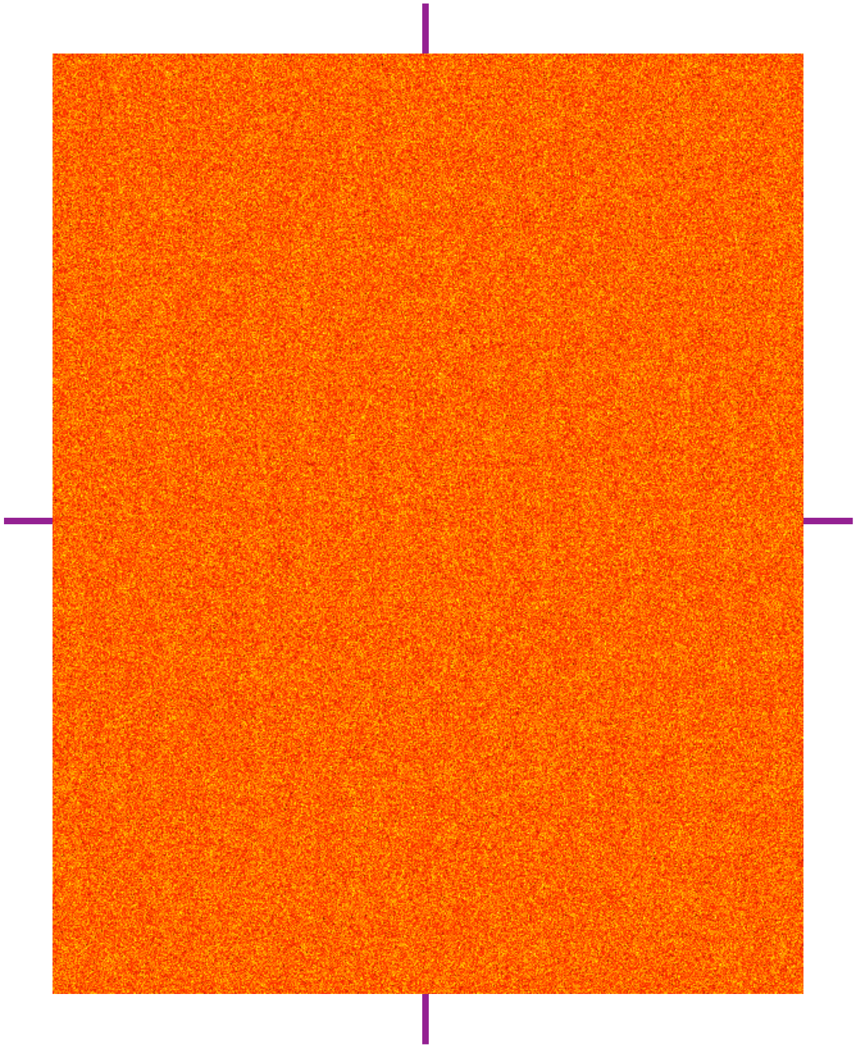
6. Use Parseval's relation for dyadic wavelets to compute $T(n)$

7. Apply steps 3-6 to evaluate $\overline{T}$ and normalize detection statistic

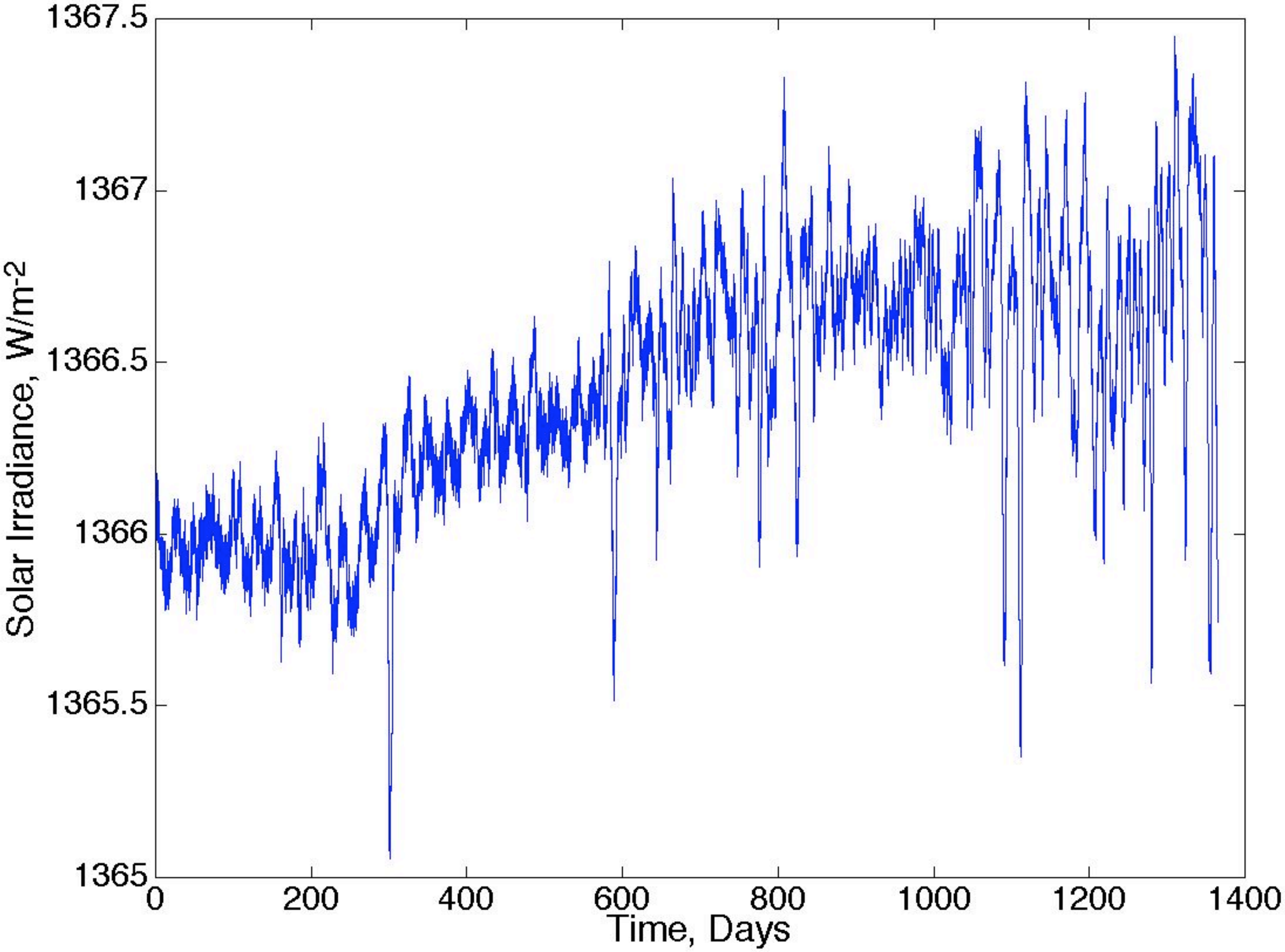
A Wavelet-Based Matched Filter (graphical version)



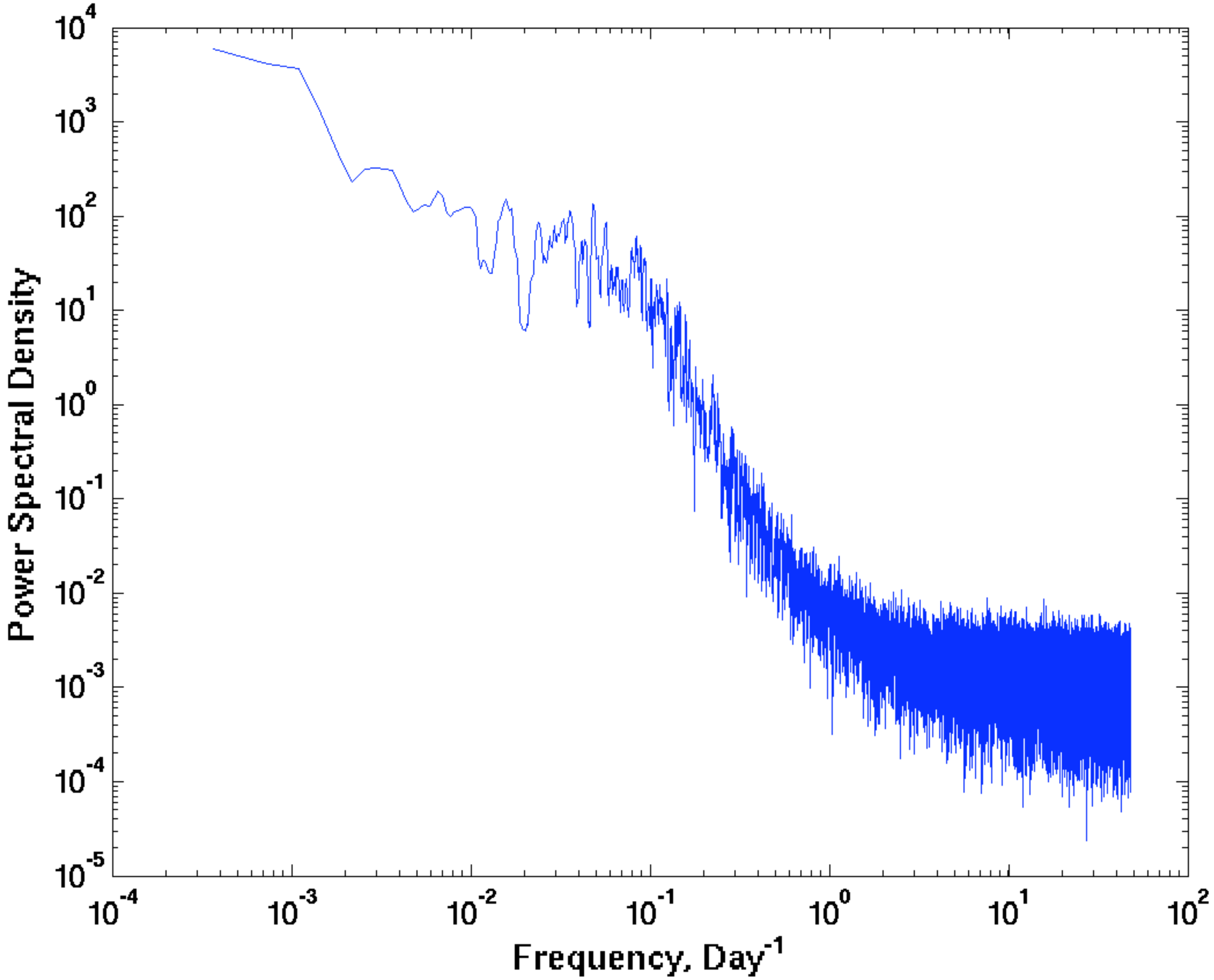
Detecting Astronomers



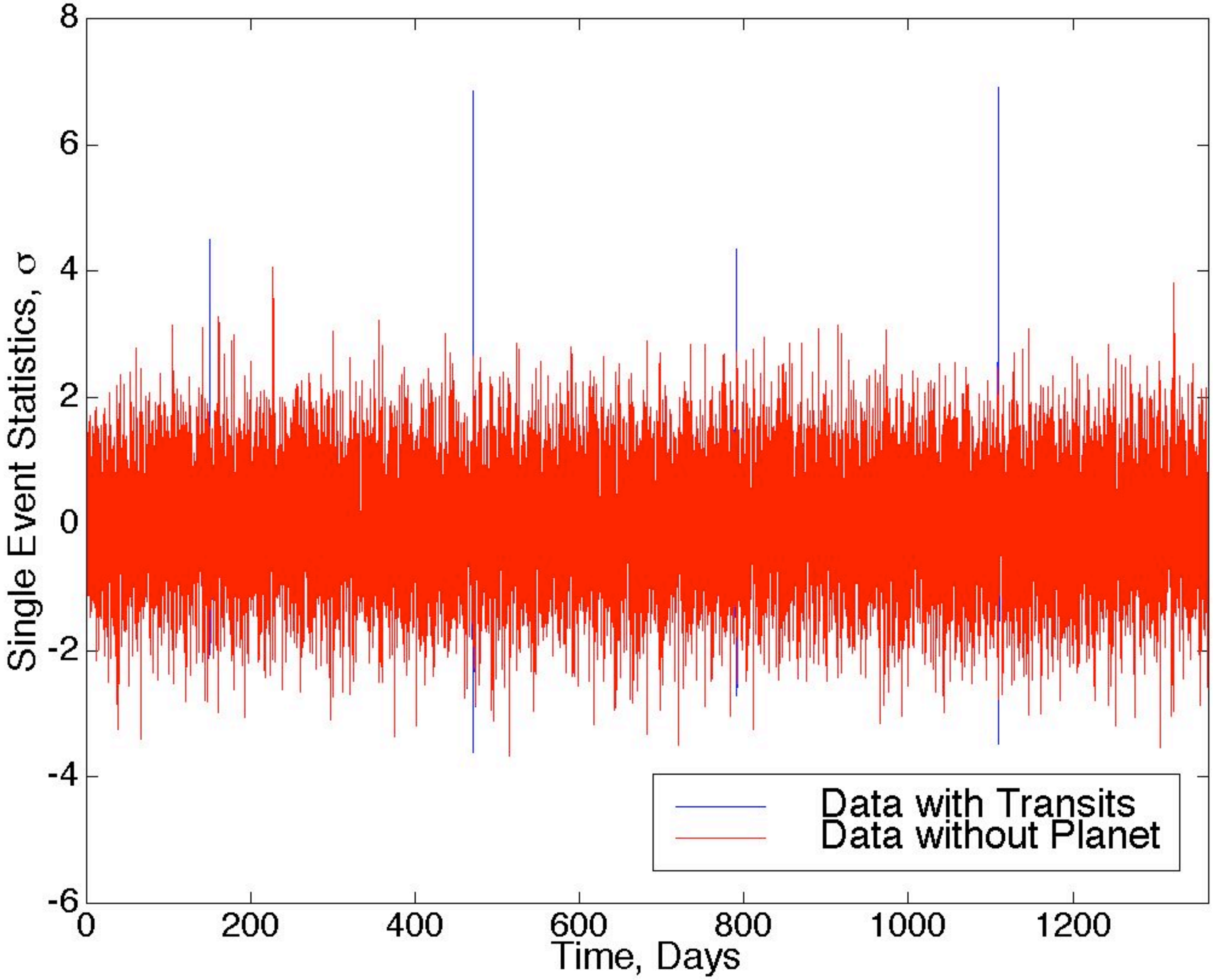
Kepler-like Noise + Transits



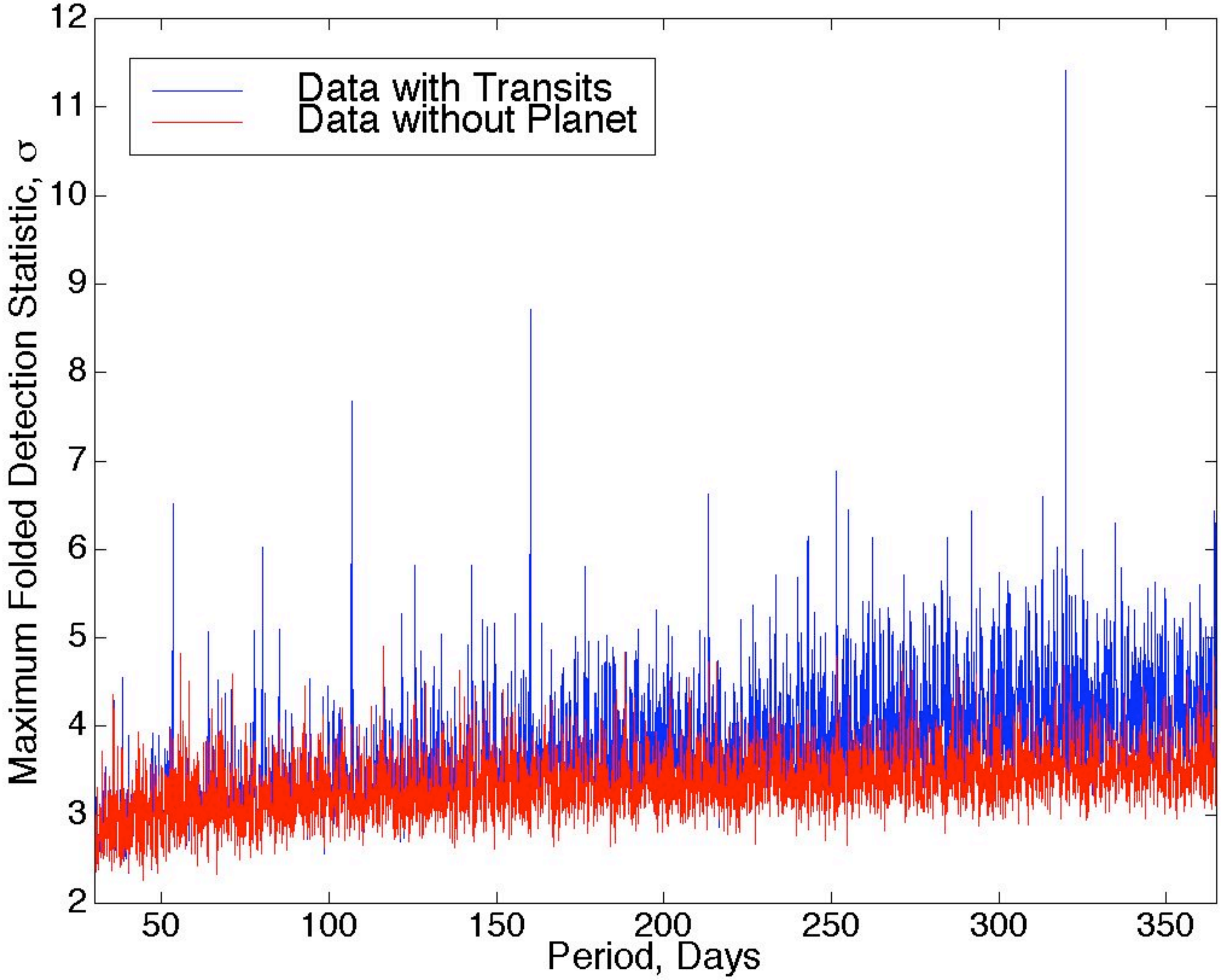
Stellar Variability



Single Transit Statistics

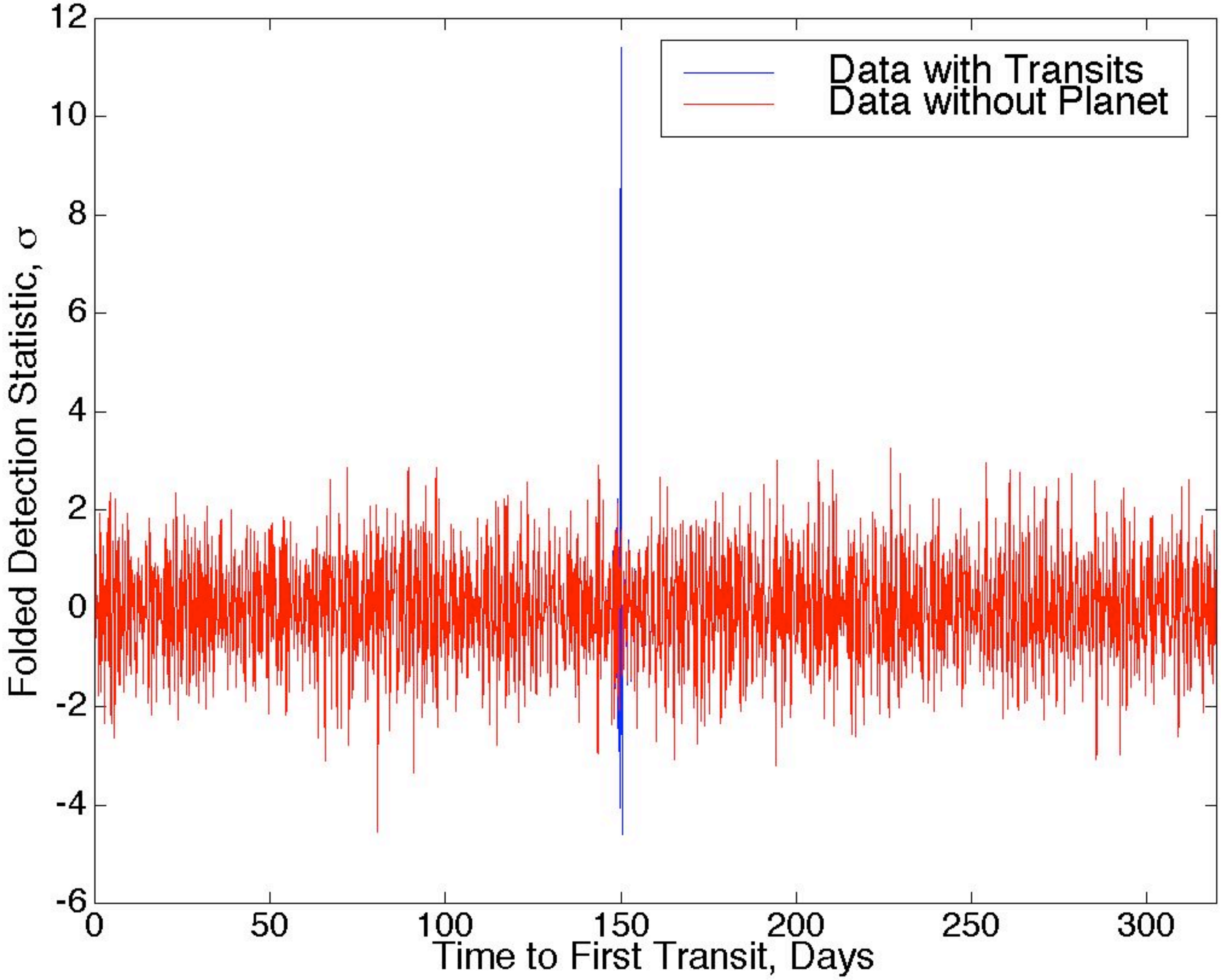


Folded Statistics



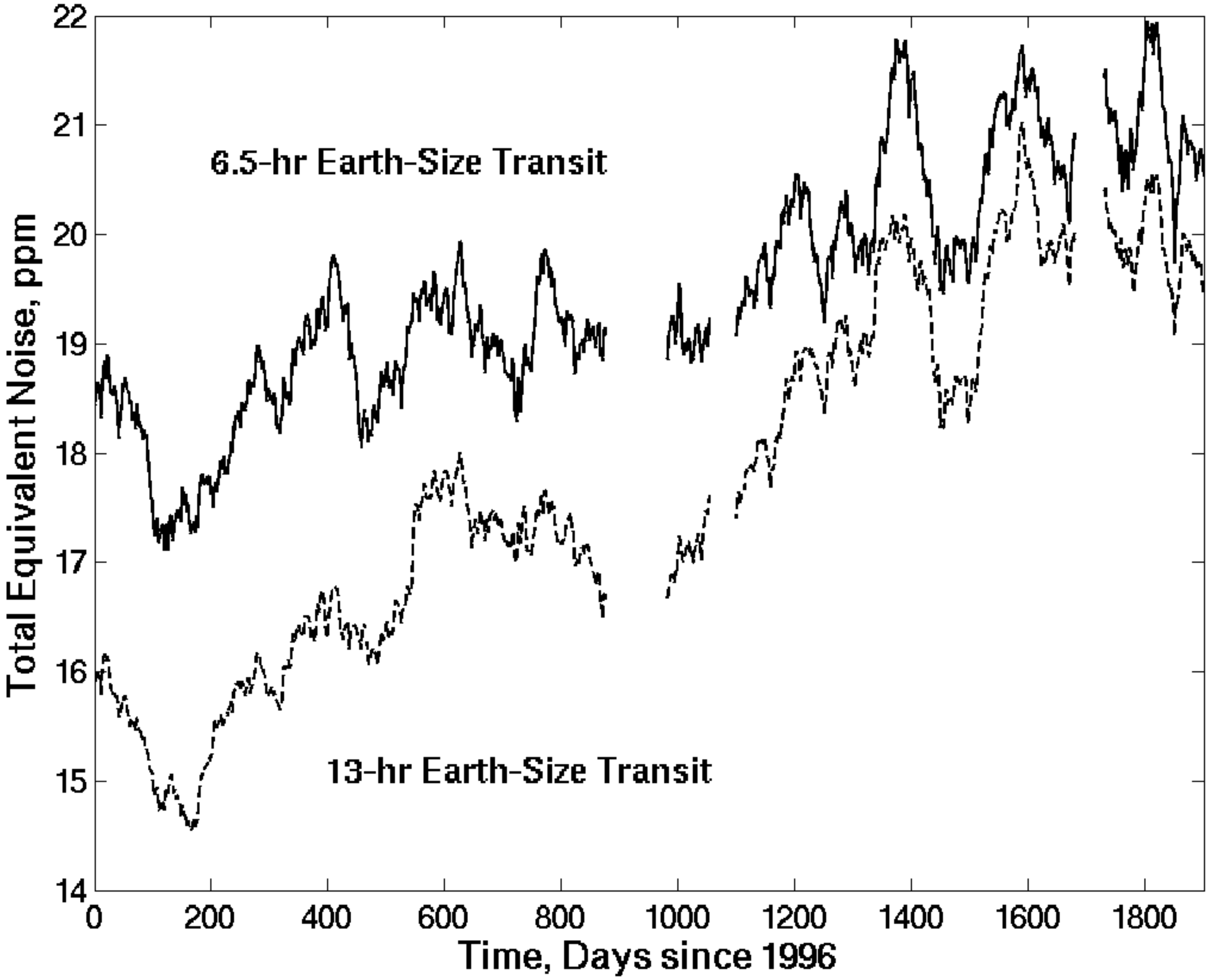
Folded Statistics for Best-Matched Period

A Search for Habitable Planets



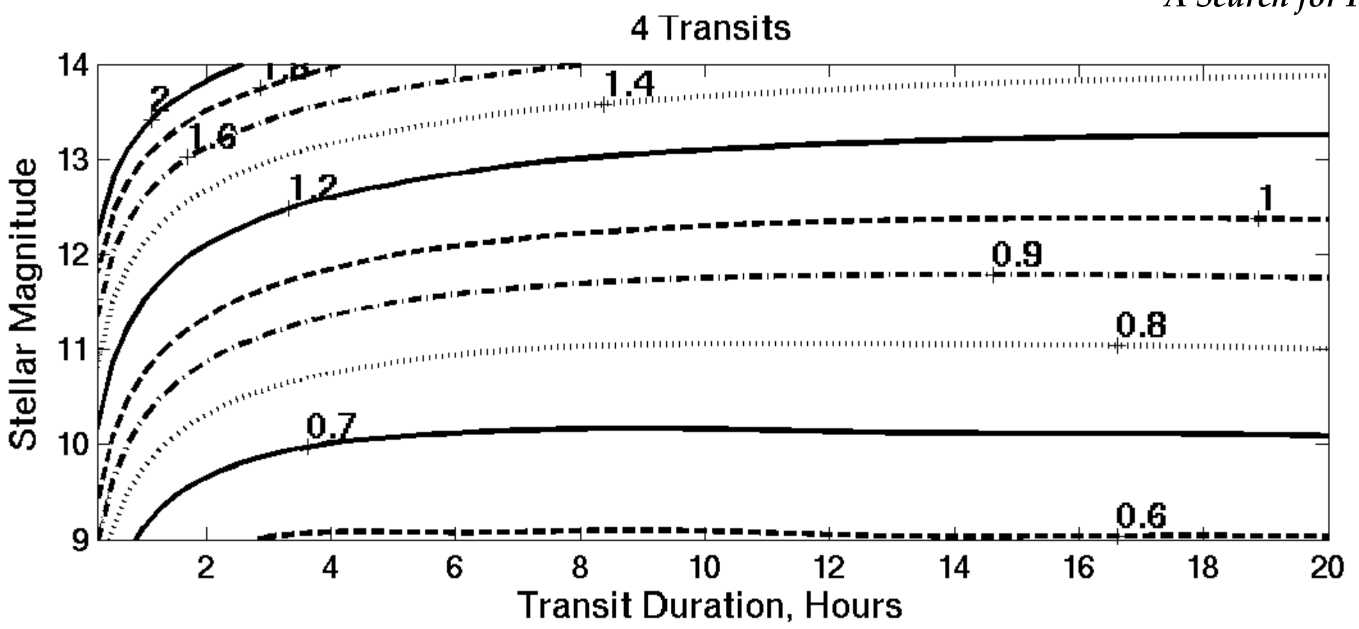
Effective Noise for Detecting Transits

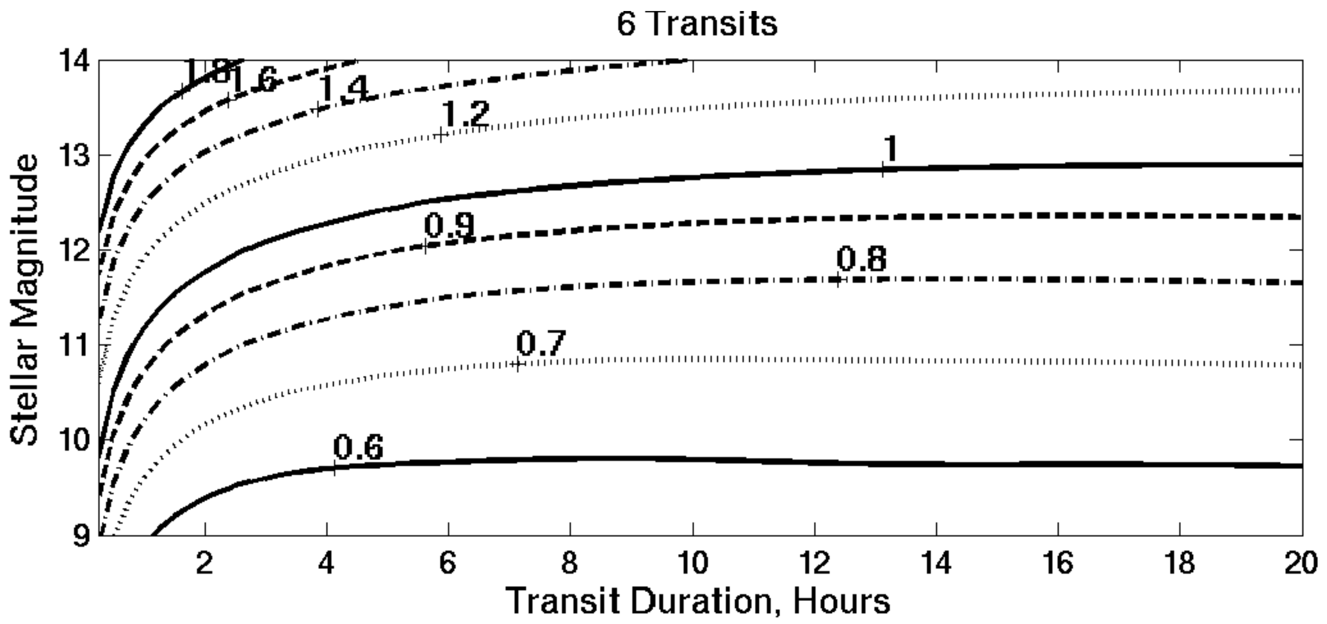
A Search for Habitable Planets



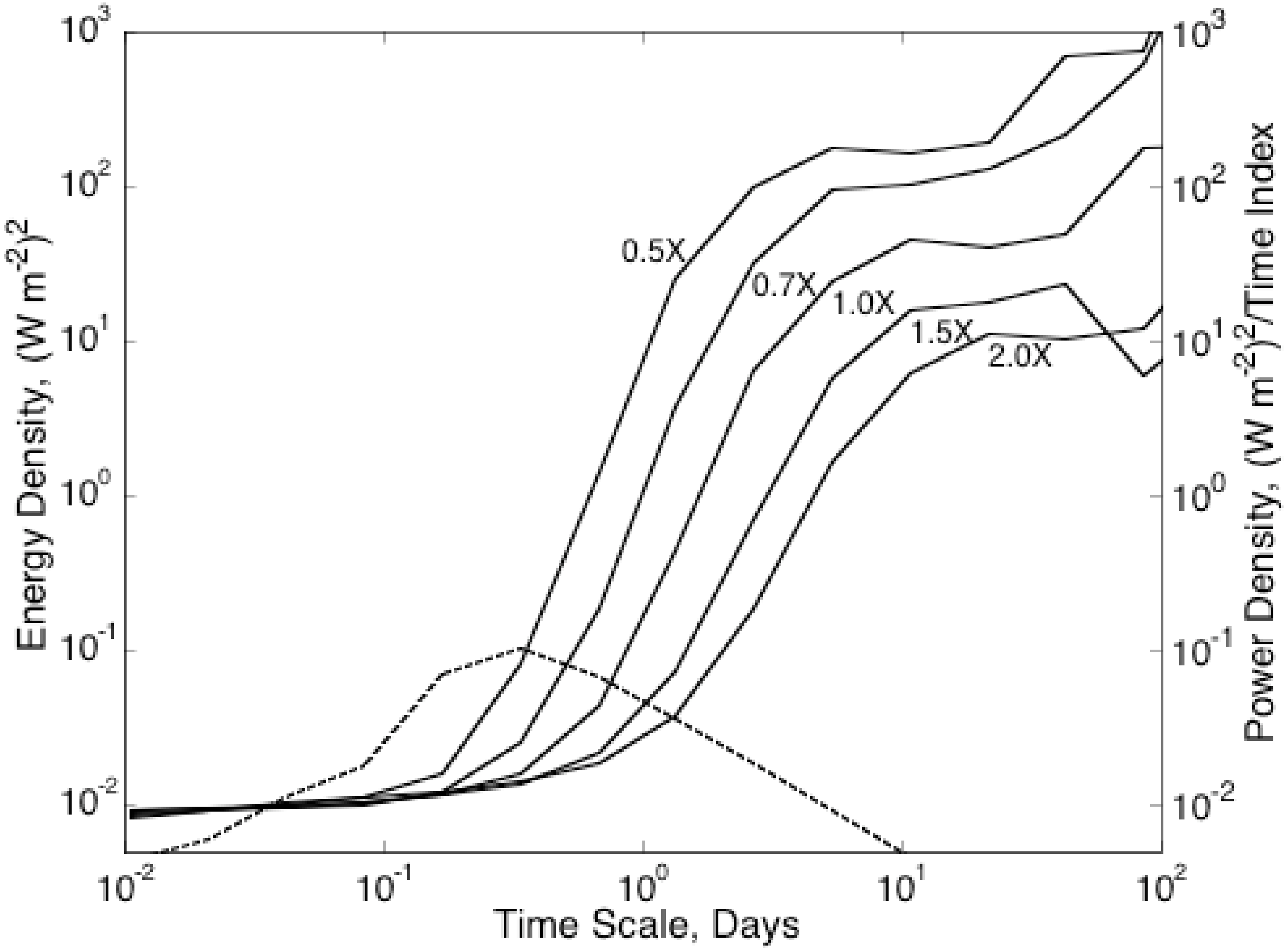
Minimum Detectable Planetary Radius

A Search for Habitable Planets


4 Transits

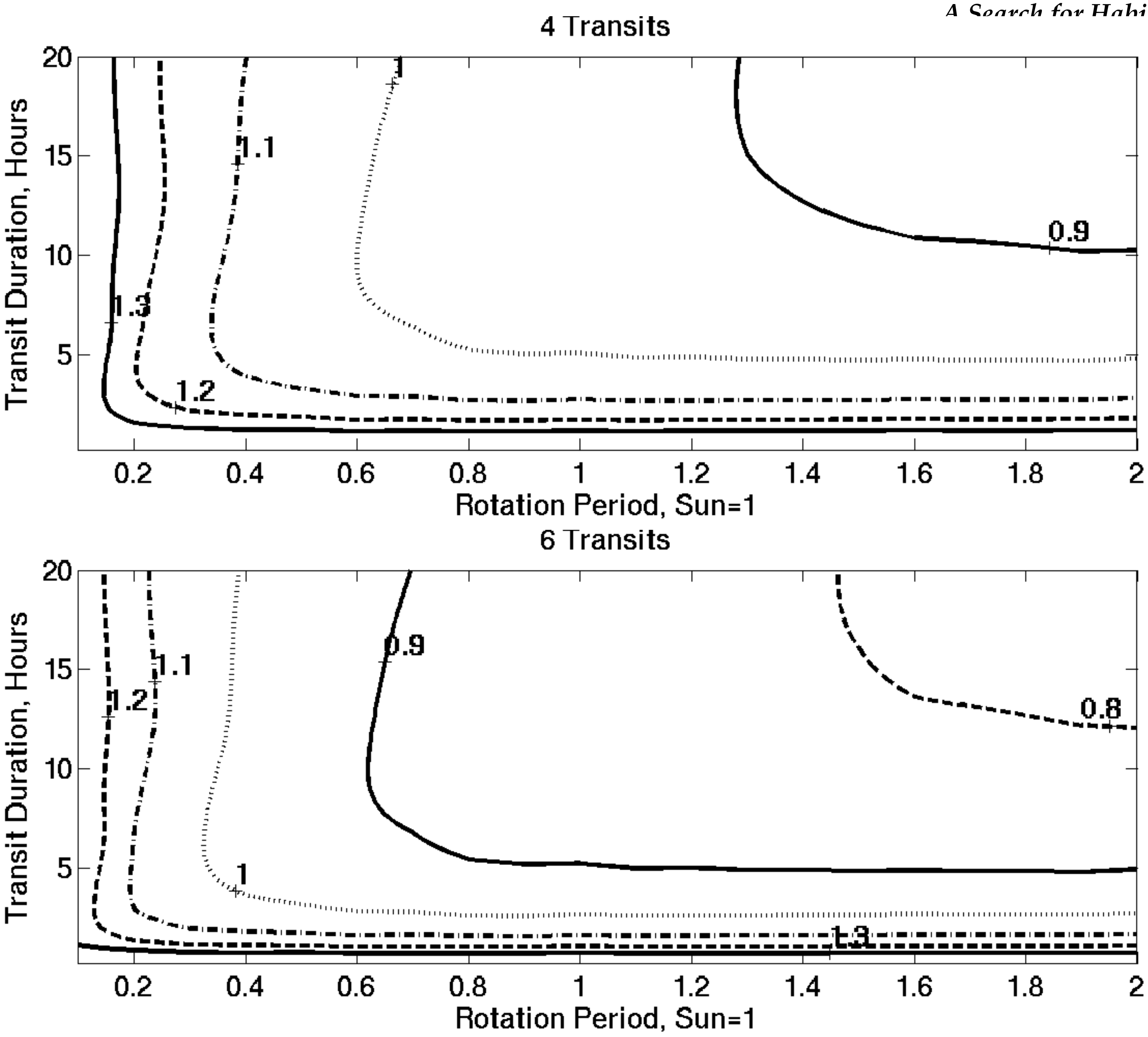

6 Transits

Stellar Variability vs. Stellar Rotation



Min. Detectable R_p vs. Stellar Rotation

A Search for Habitable Planets


4 Transits

6 Transits

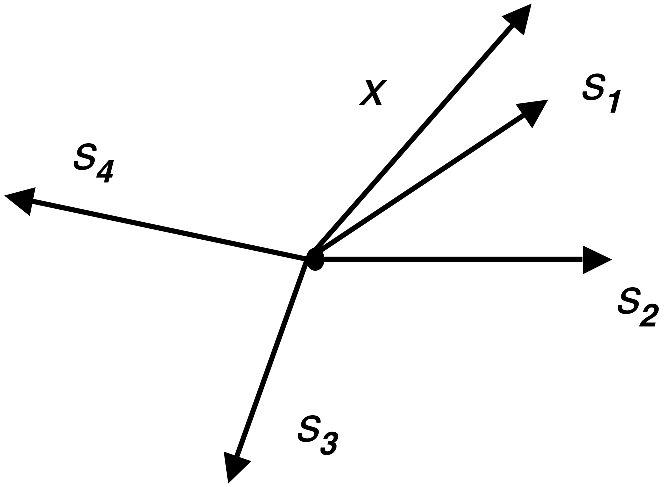
Setting the Threshold

How many statistically independent tests are we conducting?

The answer depends on the duration of the transits and range of orbital periods we are interested in, and the sampling of the observations

The search consists of:

1. Computing single transit statistics
2. Folding single transit statistics
3. Examining maximum folded statistic



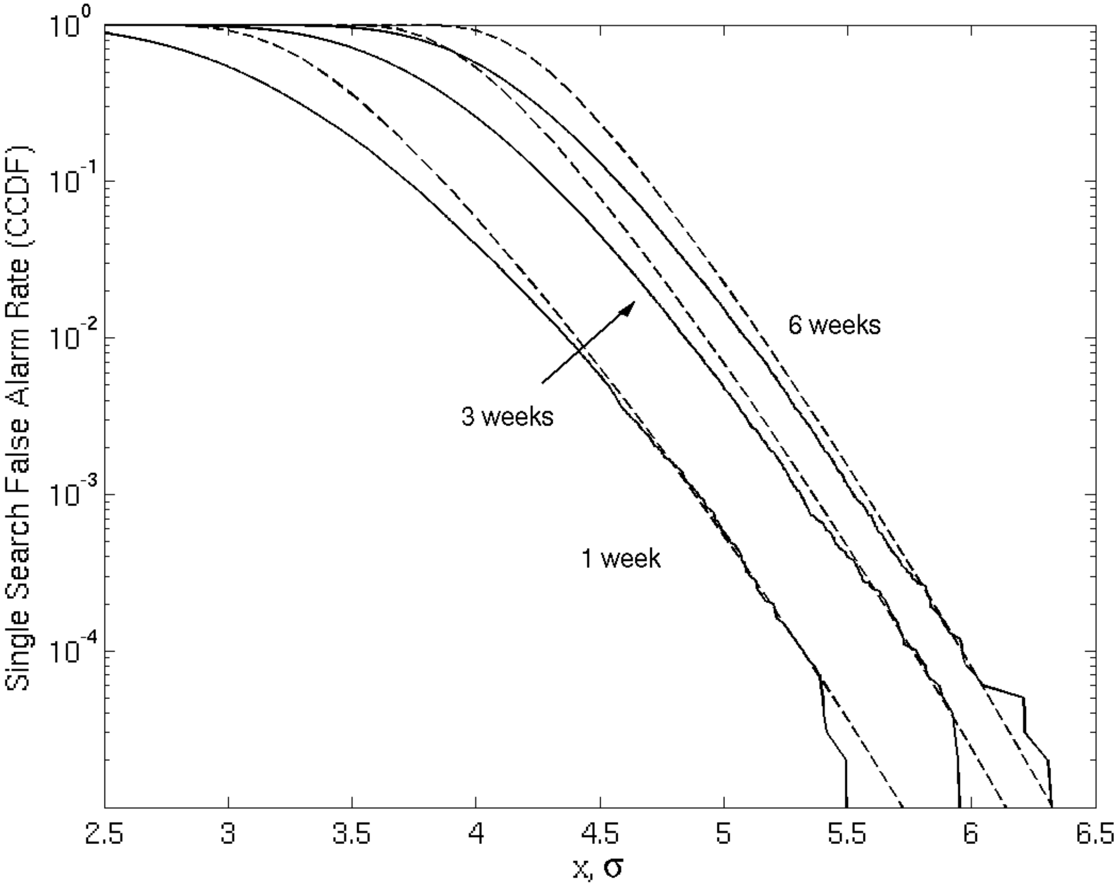
The required threshold depends on the distribution of the maximum folded statistics

Modeling the Single Search FAR

A Search for Habitable Planets

We can model the distribution of T_{max} as the max of N_{EIT} draws from WGN

$$\overline{F}(x) = 1 - G(x)^{N_{EIT}} = 1 - \left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\tfrac{1}{2} y^2\right) dy \right]^{N_{EIT}}$$



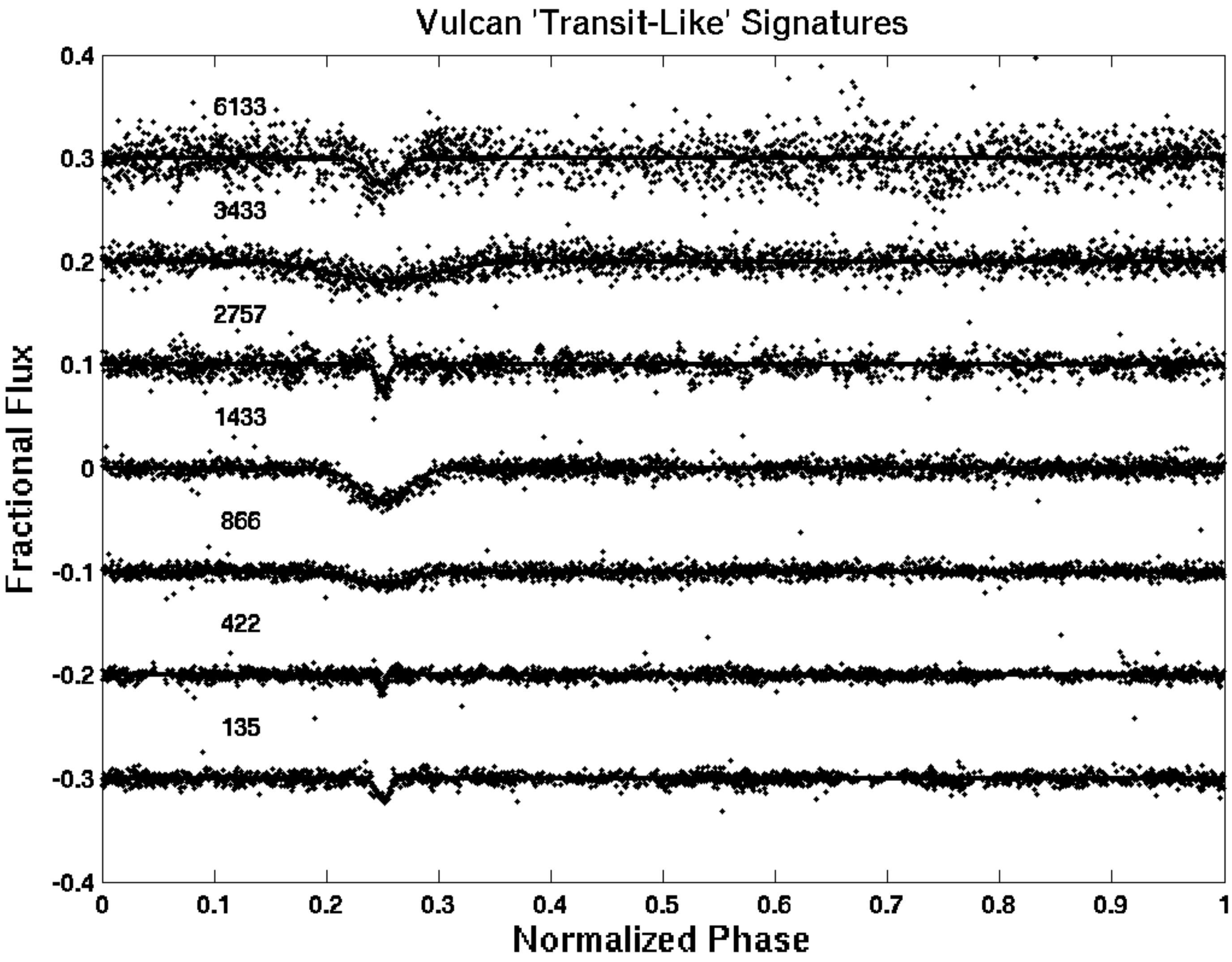
Bootstrap for Non-IID Points

Algorithm:

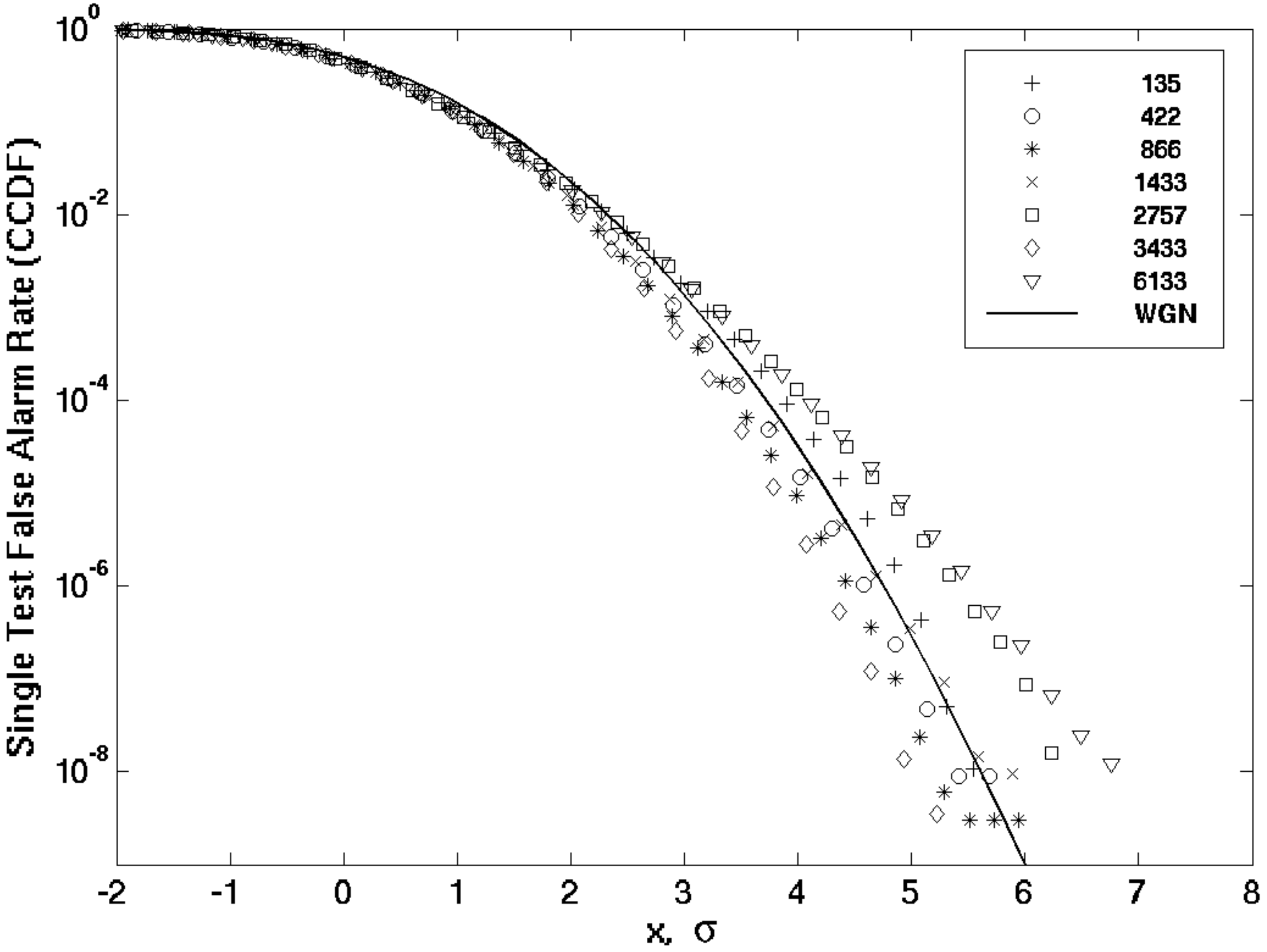
1. **Form single transit statistics from light curve with “transits” removed**
2. **Draw $N_{transit}$ statistics at random from set formed in Step 1**
3. **Form total detection statistic**
4. **Repeat steps 1-3 a large number of times**

The confidence is the fraction of times that the synthetic statistics exceed the observed statistic

Bootstrap II : Vulcan Data

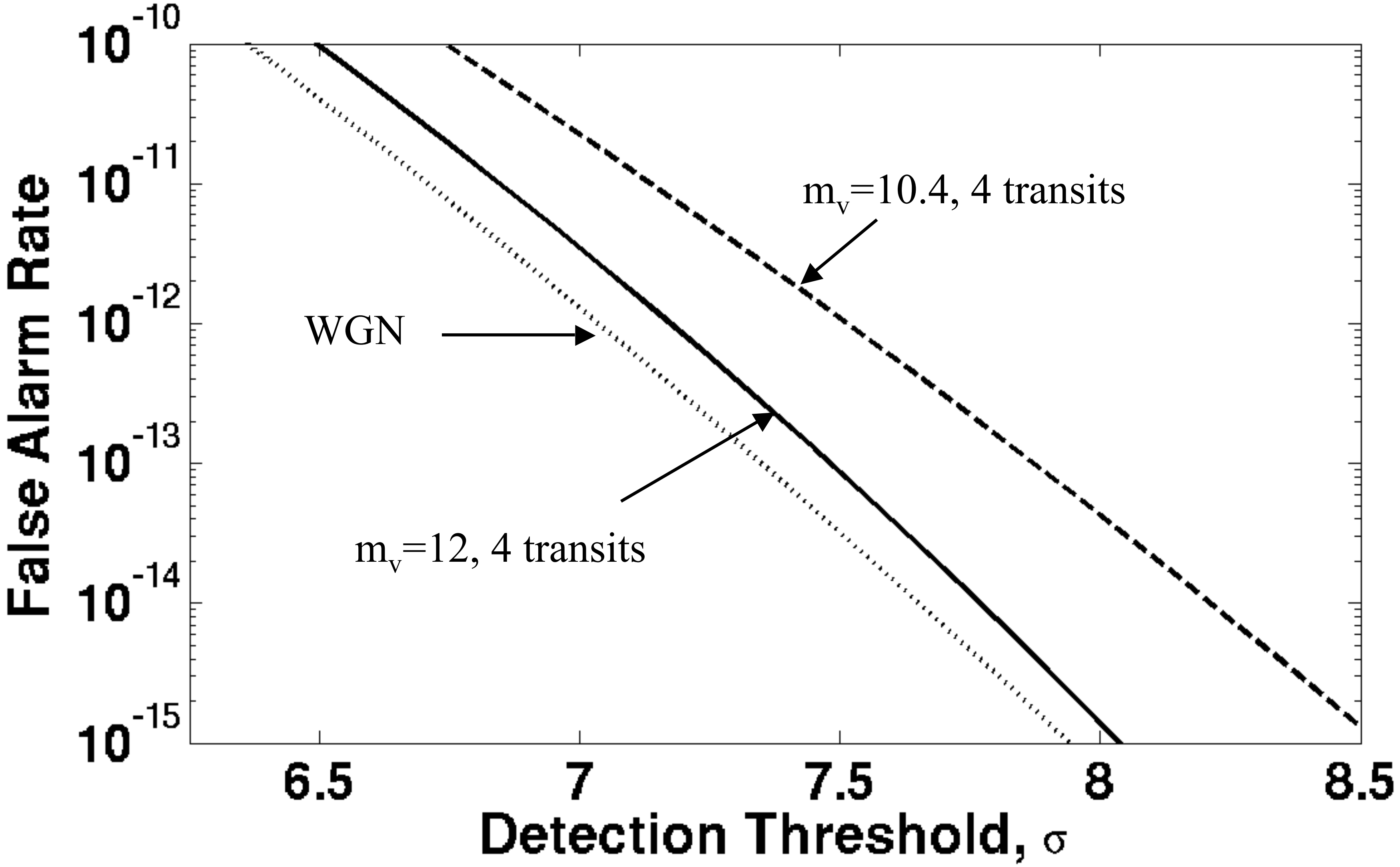


Bootstrap II : Vulcan Data



How Gaussian is Solar-Like Variability?

A Search for Habitable Planets

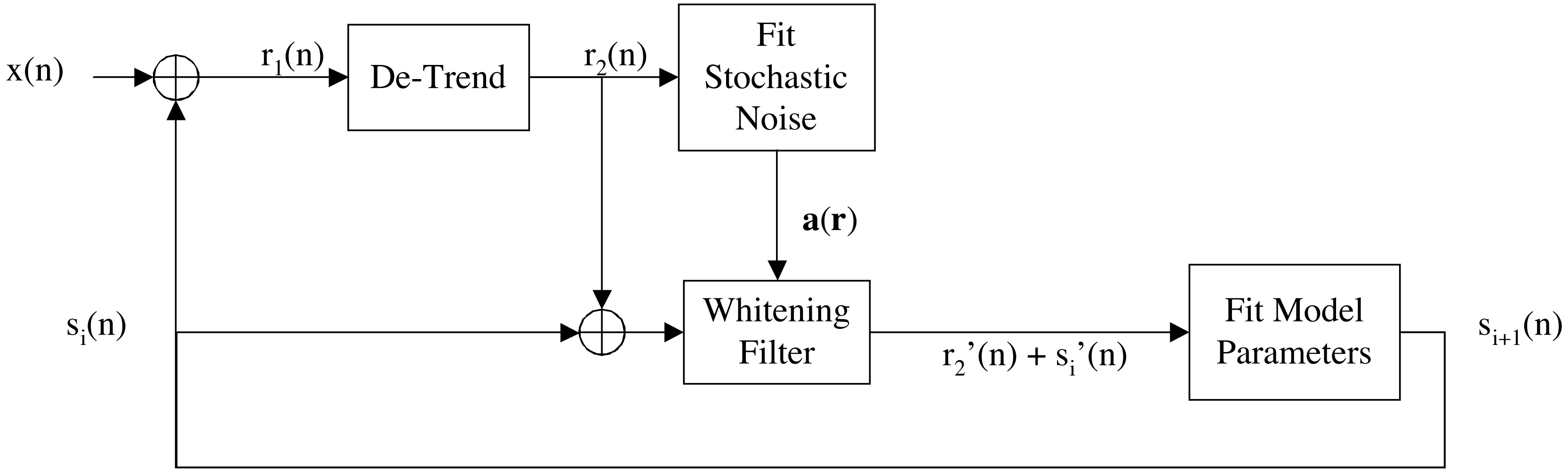


Detecting Multiple Transiting Planets

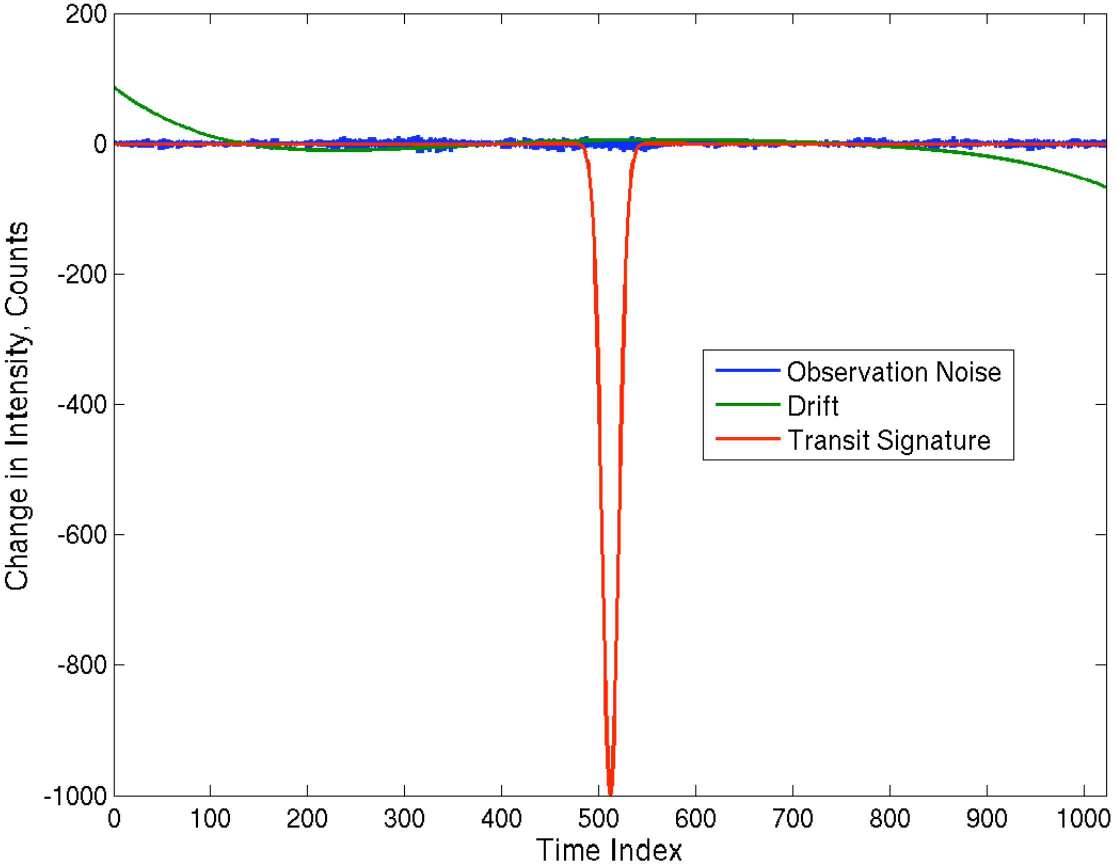
Iterative joint estimation of the planet (eclipsing binary) model parameters/characterization of the observation noise

At each iteration:

- **Subtract the current estimate of the planetary signature, $s_i(n)$, from the data, x(n)**
- **Update the noise parameters, a(r), based on analyzing the residuals r(n)**
- **Update the planetary model parameters (accounting for the whitening filter effects)**
- **Repeat until convergence is achieved**



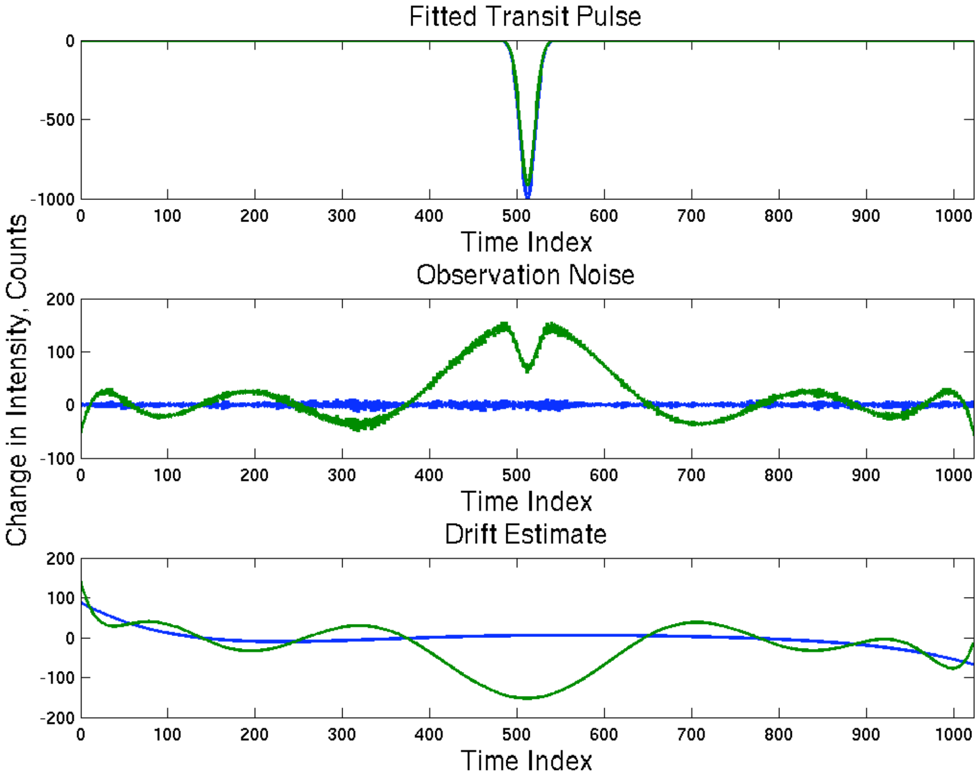
A Numerical Example



A strong, transit-like pulse (red) is added to non-white observation noise (blue) and a long term photometric drift signature (green) to generate a synthetic time series

The algorithm iteratively recovers the physical parameters of the “transiting planet” (orbital period, transit depth, transit width)

Results of Iteration 1 of MTPS

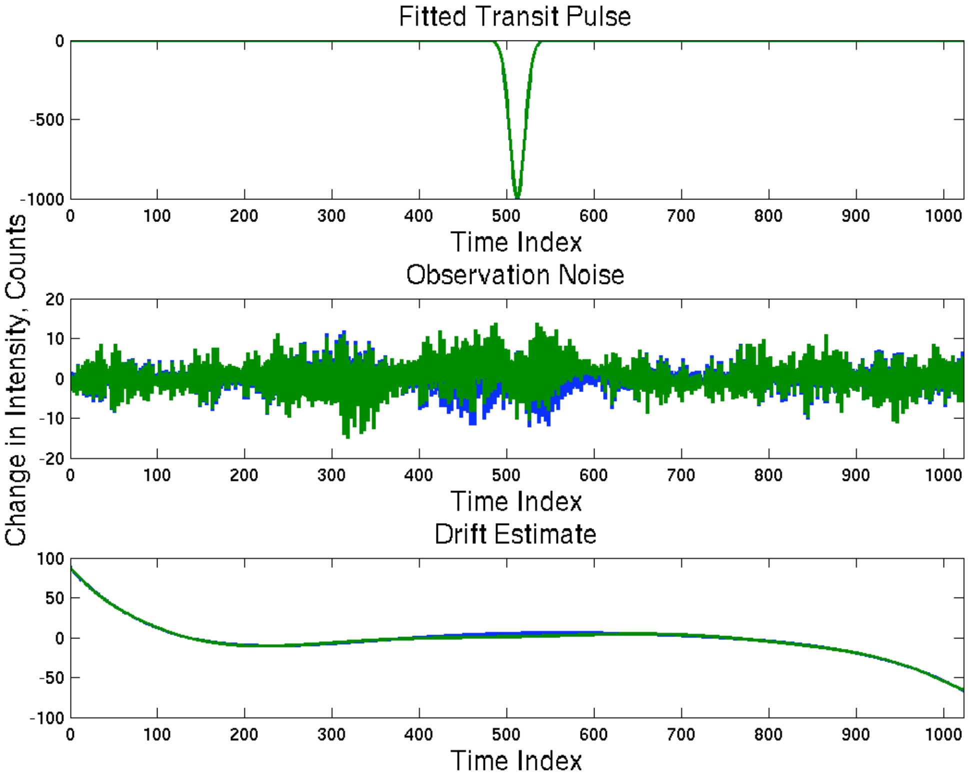


The top panel shows the fitted transit pulse (blue) and the original transit pulse (green)

The middle panel shows the reconstructed observation noise (blue) and the original observation noise (green)

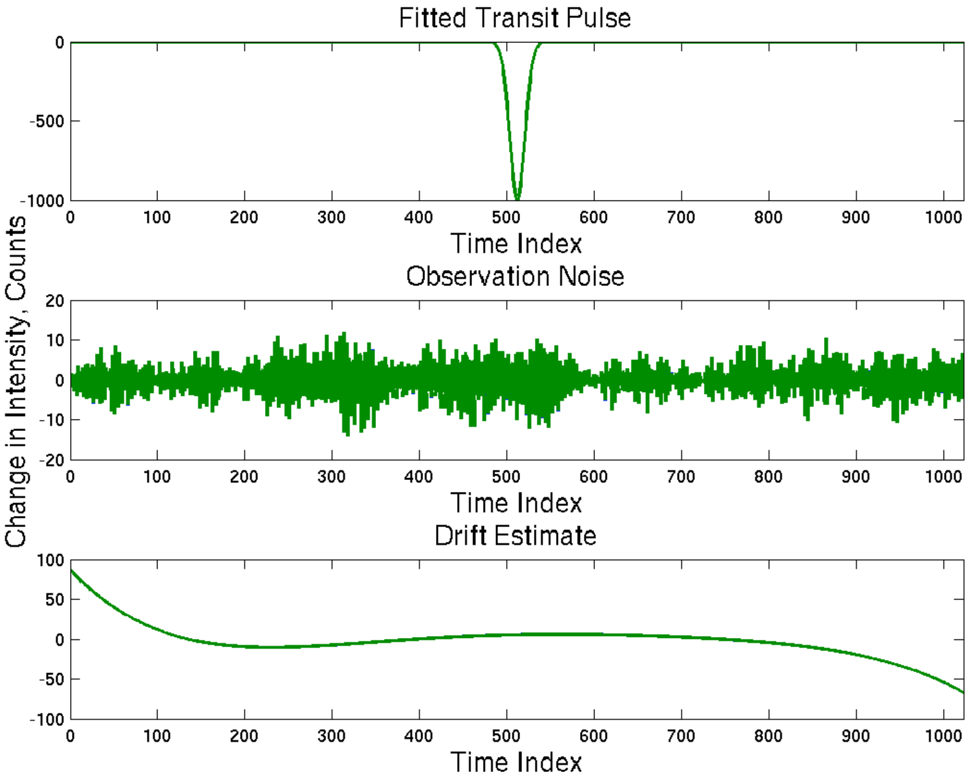
The bottom panel shows the reconstructed drift (blue) and the original drift (green)

Results of Iteration 3 of MTPS



The reconstructed signal components (transit, noise, drift) are converging to the actual components as the noise characterization improves

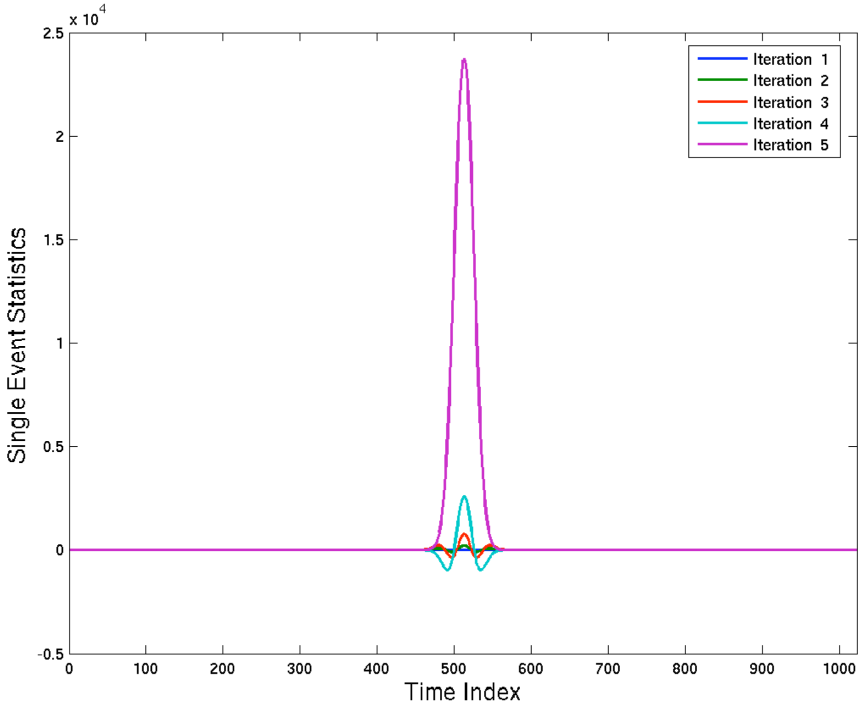
Results of Iteration 5 of MTPS



The reconstructed signal components (transit, noise, drift) have converged to the actual components (to within the limits of the stochastic noise

The apparent transit S/N has dramatically improved

The “Weak” Strong Transit Signal



The performance of detection algorithms designed to find weak signals in noisy data often degrades for strong signals

We characterize the noise for the purpose of designing an appropriate whitening filter using the observed data

Strong transit-like features corrupt the noise characterization, and the resulting whitening filter tends to annihilate the transit-like features

The MTPS algorithm iteratively improves the noise characterization by removing successively better estimates of the planetary signature from the data prior to noise characterization

Propagation of Errors

To estimate the uncertainties in the fitted planetary parameters we apply the standard propagation of errors in the whitened data space and transform these to the observation domain.

Take $\mathbf{H}$ to be the whitening filter acting on data *y* to yield whitened data $\tilde{\mathbf{y}} = \mathbf{H}\mathbf{y}$

Then the covariance matrix for $\tilde{\mathbf{y}}$ is simply $\mathbf{C}_{\tilde{y}} = \mathbf{I}$.

And let $\hat{\mathbf{y}} = \mathbf{A}\,\mathbf{x}$

be a linearized model for the data (planetary signature) with parameters $\mathbf{x}$.

Then
$$\hat{\tilde{\mathbf{y}}} = \mathbf{H}\mathbf{A}\,\mathbf{x} = \tilde{\mathbf{A}}\,\mathbf{x}$$

and
$$\mathbf{x} = \left(\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}\right)^{-1}\tilde{\mathbf{A}}^T\,\tilde{\mathbf{y}}$$

so that
$$\begin{aligned}\mathbf{C}_\mathbf{x} &= \left(\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}\right)^{-1}\tilde{\mathbf{A}}^T\left\langle\tilde{\mathbf{y}}\cdot\tilde{\mathbf{y}}^T\right\rangle\left[\left(\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}\right)^{-1}\tilde{\mathbf{A}}^T\right]^T\\ &= \left(\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}\right)^{-1}\tilde{\mathbf{A}}^T\mathbf{I}\tilde{\mathbf{A}}\left(\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}\right)^{-1} = \left(\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}\right)^{-1}\\ &= \left(\mathbf{A}^T\mathbf{H}^T\mathbf{H}\mathbf{A}\right)^{-1}\end{aligned}$$

Note that the actual calculation can be performed by applying the whitening transform to each component vector in the linearized planetary model.

False Positive Rejection: Centroids

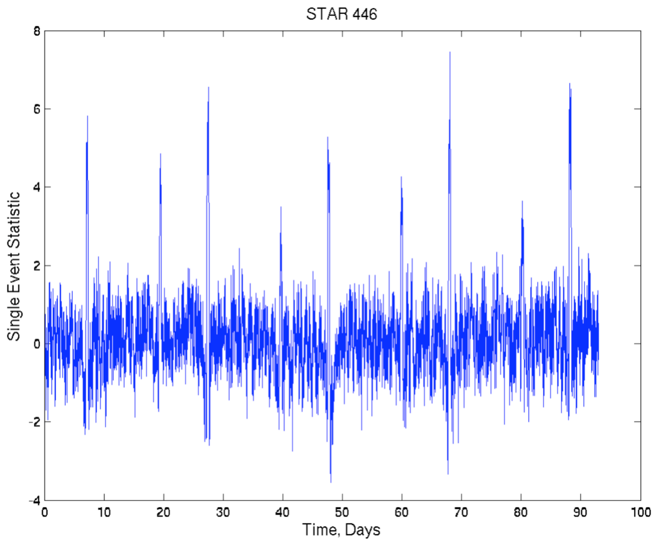
Given: background contaminating source (eclipsing binary)

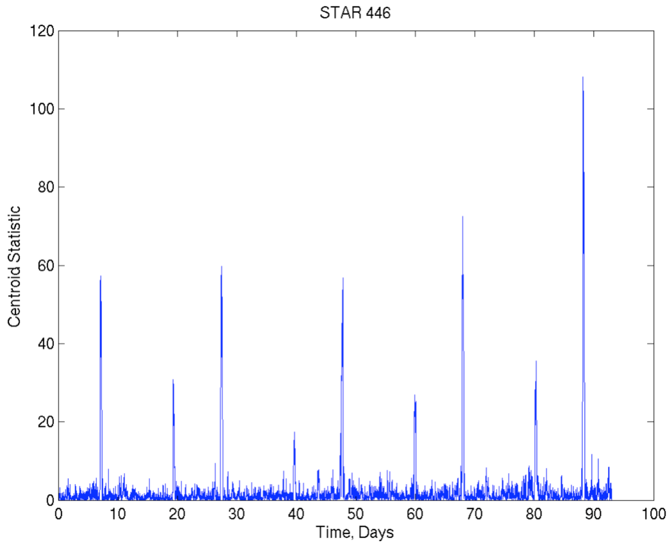
Offset from target with flux B by ΔX

Observed centroid: $\delta x = \delta b \, \Delta X/(B+b)$

b<<B, so $\delta x / \Delta X = \delta b / B$

Background sources of confusion can be rejected if there is a significant correlation between the observed flux variations and the observed centroid variations





SUMMARY

Searching for Transiting Planets Requires:

A matched filter (or equivalent) detector is optimal

We’ve presented a numerically efficient, optimal, adaptive wavelet-based matched filter

Careful consideration of search parameters

Establishing a rational Threshold

A method for dealing with strong transit-like signatures
(and permit detection of multiple transiting planets)

Ways to eliminate obvious false positives

Patience and Perseverance!