Detection of Very Short Signal Segments

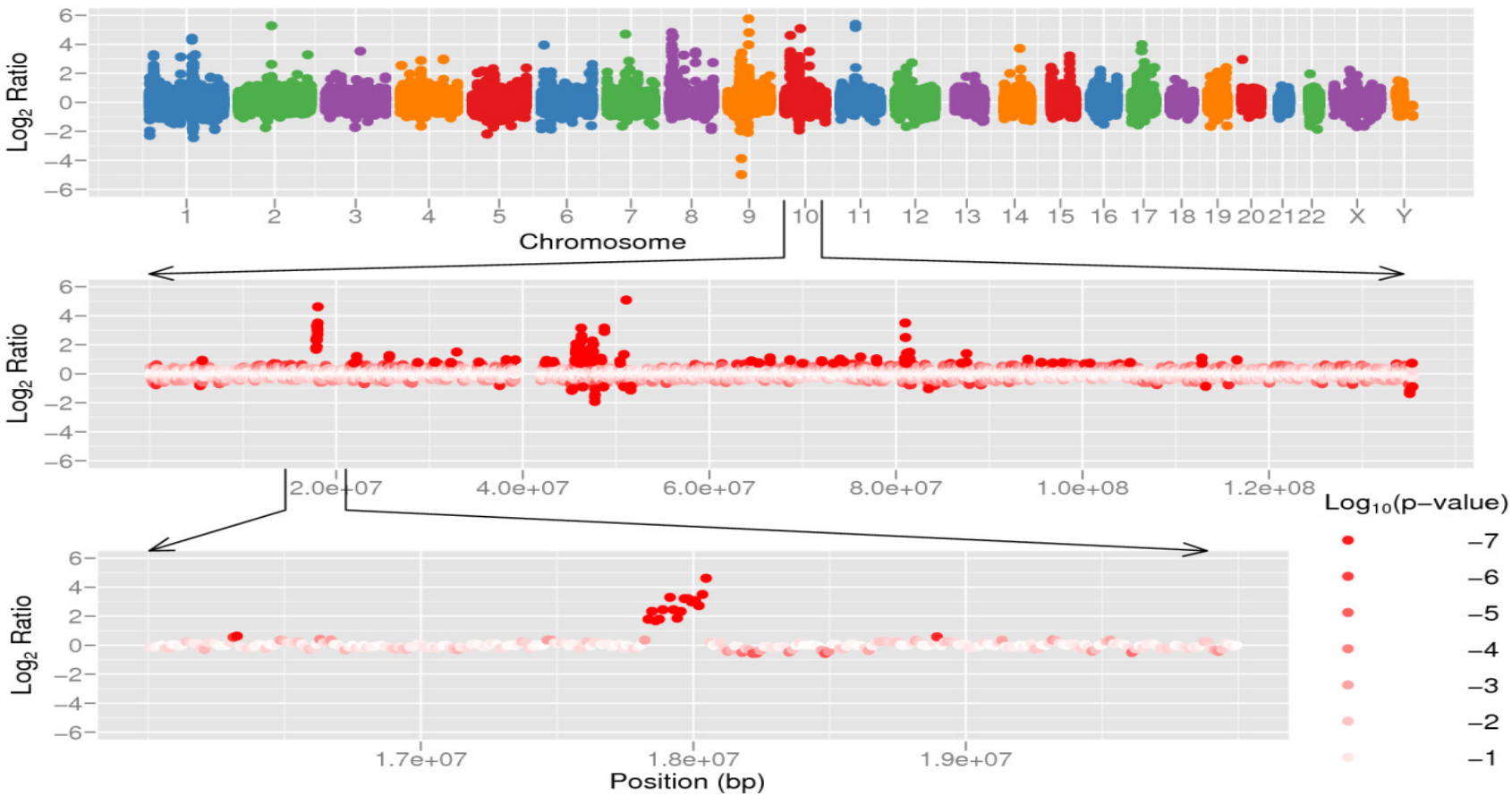
Ming Yuan

Morgridge Institute for Research
and
Department of Statistics
University of Wisconsin-Madison

myuan@stat.wisc.edu
http://www.stat.wisc.edu/~myuan

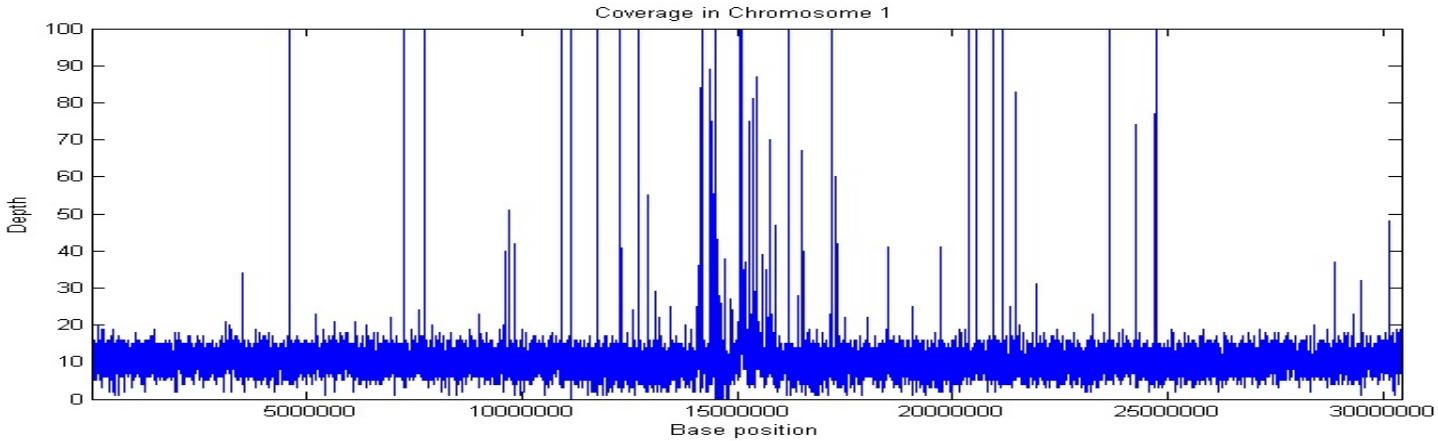
(Joint work with Tony Cai)

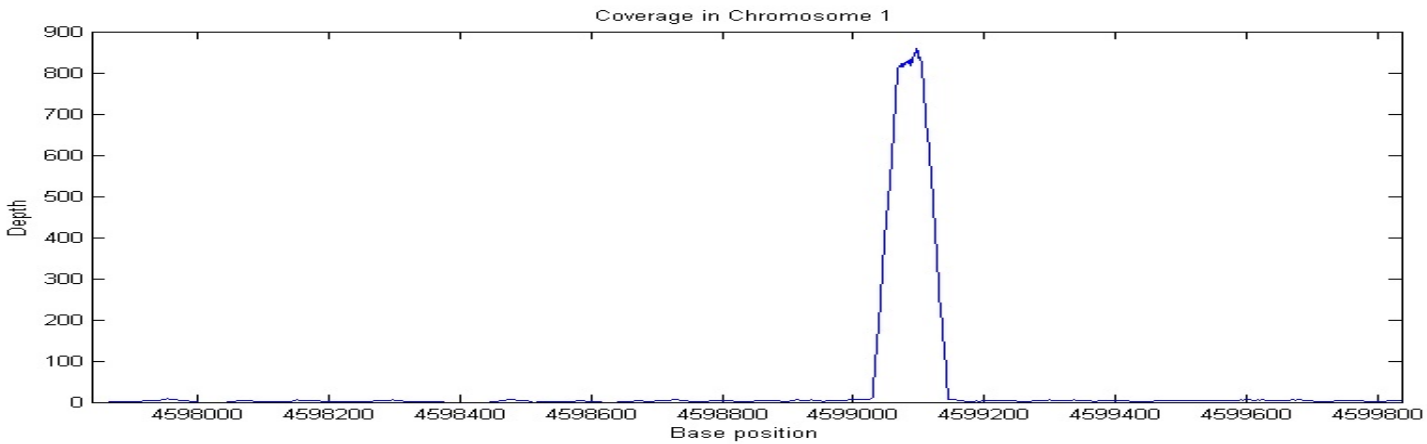
Copy Number Variation



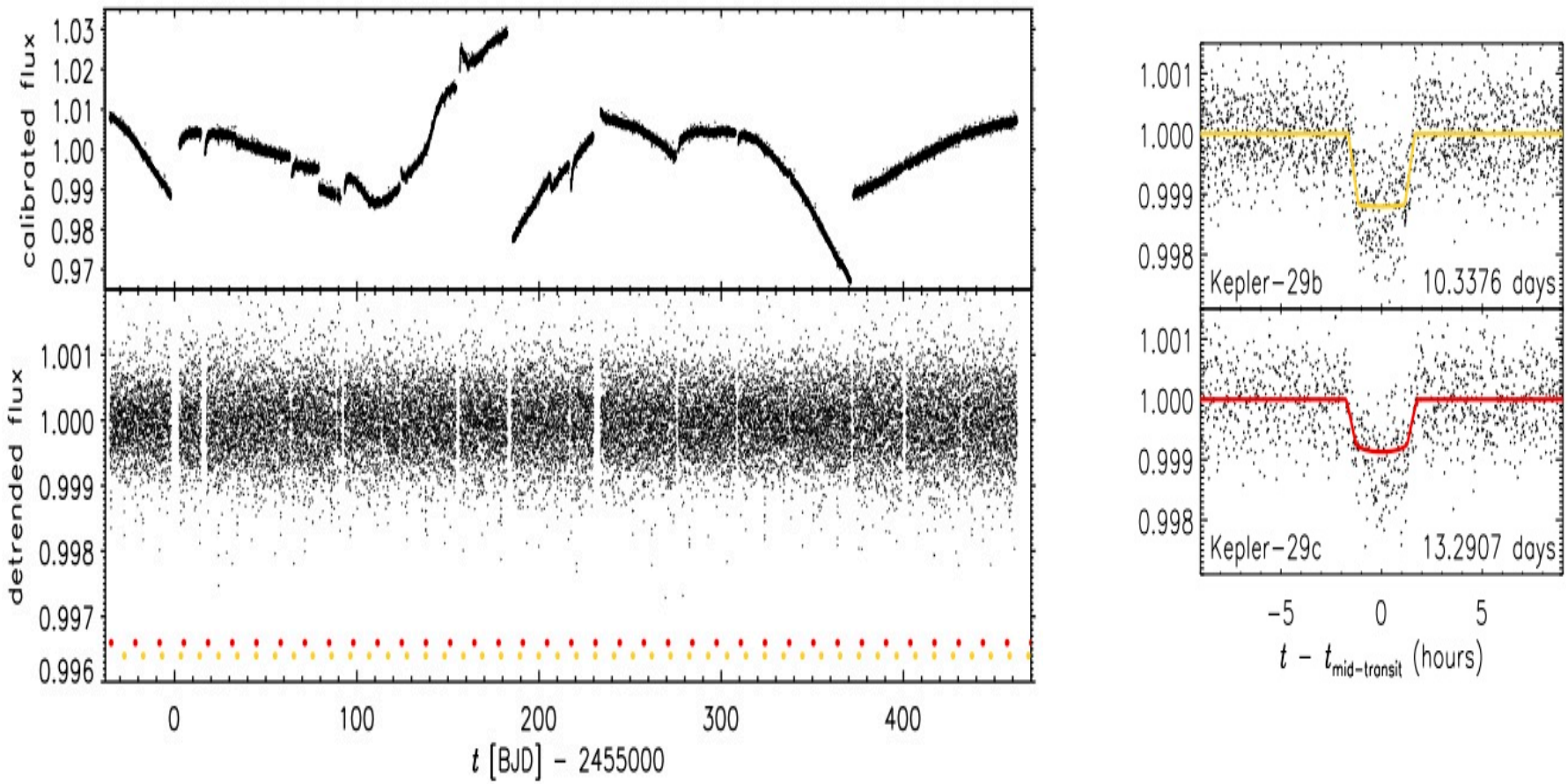
[Xie and Tammi, 2009]

PEAK DETECTION





Astronomy



[Fabrycky et al., 2012] `http://kepler.nasa.gov/Mission/discoveries/kepler29c/`

Structured Signal Detection

$$X_i = \mu_i + \varepsilon_i, \qquad i = 1, 2, \ldots, n$$

- Without signal
$$\mu_1 = \mu_2 = \ldots = \mu_n = 0$$
- With signals located at unknown segments $\mathcal{S} = \{S_j = (a_j, b_j]\}$
$$\mu_i = \begin{cases} f_j((i - a_j)/(b_j - a_j)) & \text{if } i \in S_j \\ 0 & \text{otherwise} \end{cases}$$
  - Short signal – $|S_j| < n^{\xi}$ for some $\xi < 1$
- Examples
  - Biology – Copy number variation
  - Engineering – Peak detection
  - Astronomy – Detecting planets (see, e.g., Fabrycky et al., 2012)
  - ……

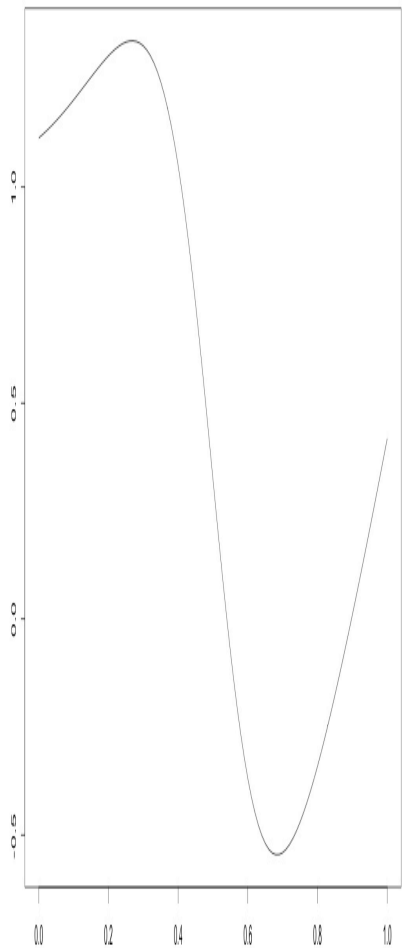
Detection of Signals

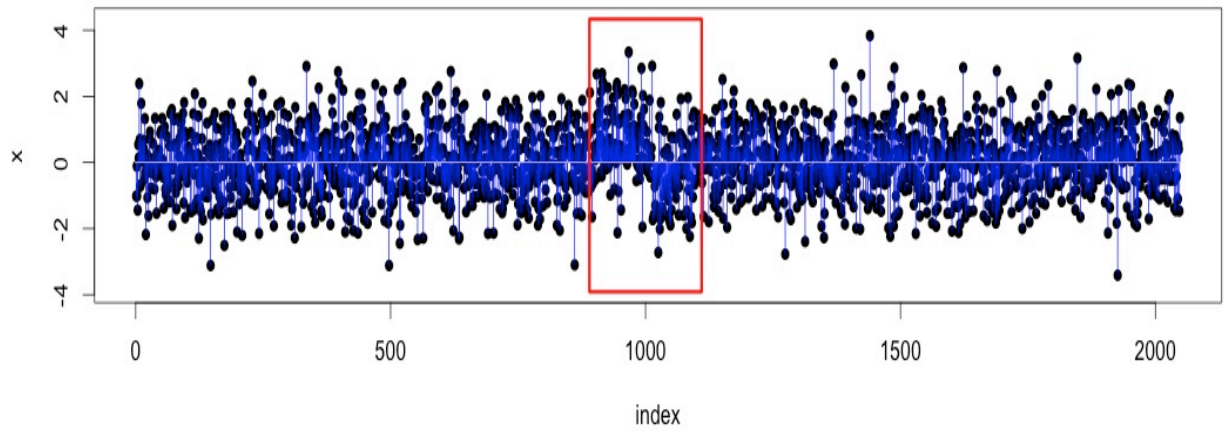
- Hypothesis testing $H_0 : \mu_i = 0$ vs

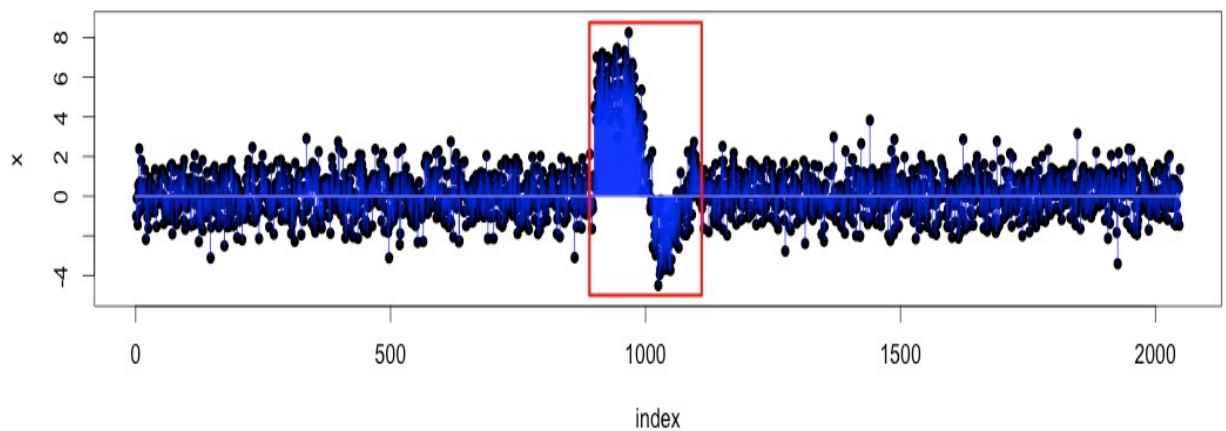
$$H_a : \mu_i = \begin{cases} f((i-a)/(b-a)) & \text{if } i \in S := (a, b] \\ 0 & \text{otherwise} \end{cases}$$

- Questions of interests
  - When is a signal detectable?
  - How to detect a "detectable" signal?
- Determining factors
  - Signal amplitude – $A = \|f\|_{L_2}$
  - Signal duration – $d = b - a$

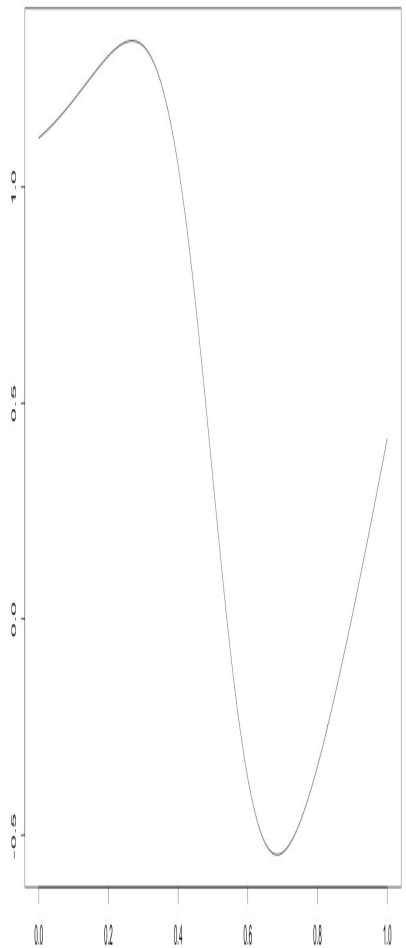
Effect of Amplitude

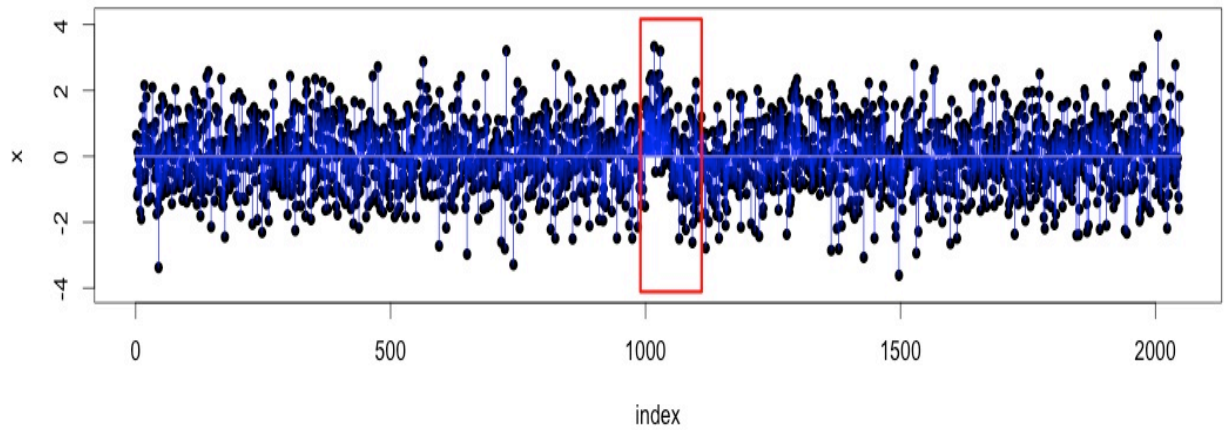


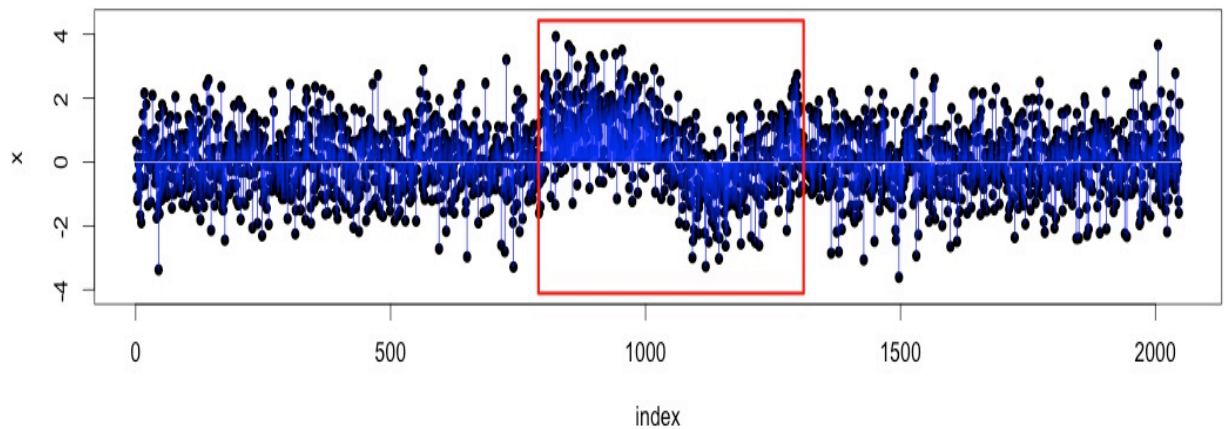




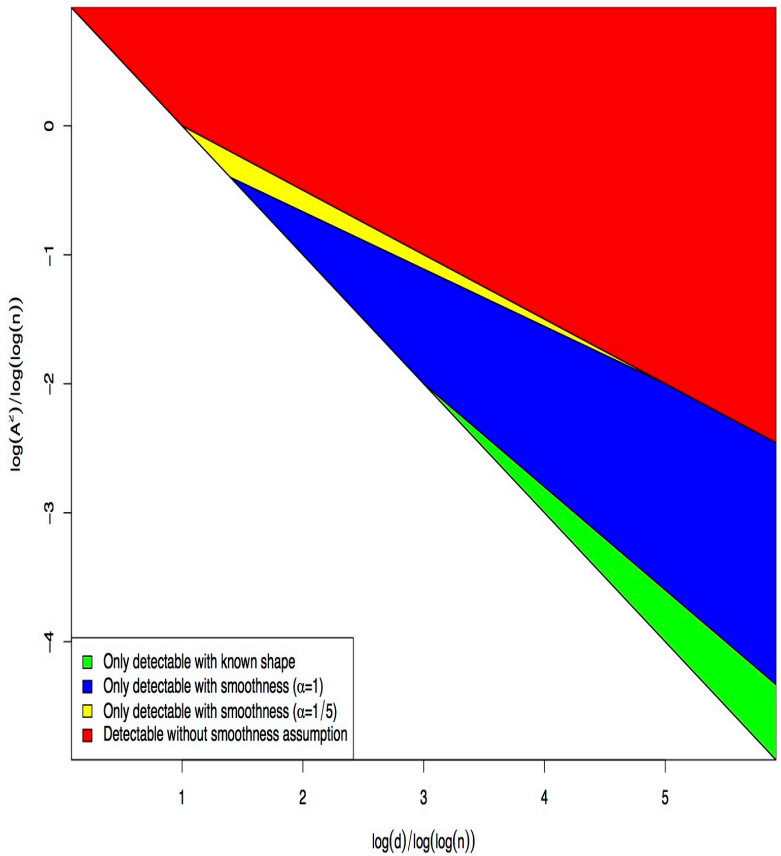
Effect of Duration







Optimal Rates of Detection



- Signal of known shape

$$A_n^2 \asymp d_n^{-1} \log n$$

- Signal of arbitrary shape

$$A_n^2 \asymp d_n^{-1} \log + \sqrt{d_n^{-1} \log n}$$

- Signal of smooth shape
  - Hölder class with $\alpha = 1$
  - $\alpha = 1/5$

Signals with Known Shape

$$f_n(\cdot) = A_n f_0(\cdot) \qquad \text{where} \qquad \int_0^1 f_0^2 = 1$$

- Constant signals – Arias-Castro et al. (2004), Jeng et al. (2004)....
- Known position and duration – Likelihood ratio type test

$$L_{(a_n, a_n+d_n]} := \left(\sum_i f_0^2(i/d_n) \right)^{-1/2} \sum_{i=1}^{d_n} X_{a_n+i} f_0\left(\frac{i}{d_n}\right)$$

- Unknown position and duration – scan statistic

$$L_n = \max_{a_n, d_n} L_{(a_n, a_n+d_n]} \Longrightarrow T_n := \begin{cases} \text{reject } H_0 & \text{if } L_n \geq c_n \\ \text{accept } H_0 & \text{otherwise} \end{cases}$$

- With critical value $c_n = c_0\sqrt{\log n}$, T_n can detect any signal such that

$$A_n^2 \gtrsim d_n^{-1} \log(n/d_n)$$

Optimality

Optimal rate of detection – $A_n^2 \asymp d_n^{-1} \log(n/d_n)$

- Detection is achieveable with $A_n^2 \gtrsim d_n^{-1} \log(n/d_n)$

$$\mathbb{P}(T_n = 1|H_0) + \sup_{H_1} \mathbb{P}(T_n = 0|H_1) \to 0.$$

- For signals with $A_n^2 \lesssim d_n^{-1} \log(n/d_n)$, any test is powerless in that

$$\inf_{\tilde{T}} \left\{ \mathbb{P}(\tilde{T} = 1|H_0) + \sup_{H_1} \mathbb{P}(\tilde{T} = 0|H_1) \right\} \to 1.$$

- Effect of multiplicity
  - Known location – d_n^{-1}
  - Price for not knowing the location – $\log n$

Signals with Arbitrary Shape

- Evidence based on quadratic statistic

$$L_{(a_n,a_n+d_n]} := 2^{-1/2}(d_n^{1/2} + \log^{1/2} n)^{-1} \sum_{i=1}^{d_n} (X_{a_n+i}^2 - 1)$$

- Scan over all possible segments

$$L_n = \max_{a_n,d_n} L_{(a_n,a_n+d_n]} \Longrightarrow T_n := \begin{cases} \text{reject } H_0 & \text{if } L_n \geq c_n \\ \text{accept } H_0 & \text{otherwise} \end{cases}$$

- T_n ($c_n = c_0\sqrt{\log n}$) can detect any signal such that

$$A_n^2 \gtrsim \left(\frac{\log n}{d_n}\right) + \left(\frac{\log n}{d_n}\right)^{1/2}$$

  - For short signals ($d_n \ll \log n$) – $A_n^2 \gtrsim d_n^{-1} \log n$
  - For long signals ($d_n \gg \log n$) – $A_n^2 \gtrsim (d_n^{-1} \log n)^{1/2}$

Optimality

Optimal rate of detection – $A_n^2 \asymp d_n^{-1} \log(n) + (d_n^{-1} \log(n))^{1/2}$

- Every test is powerless if

$$A_n^2 \lesssim \left(\frac{\log n}{d_n}\right) + \left(\frac{\log n}{d_n}\right)^{1/2}$$

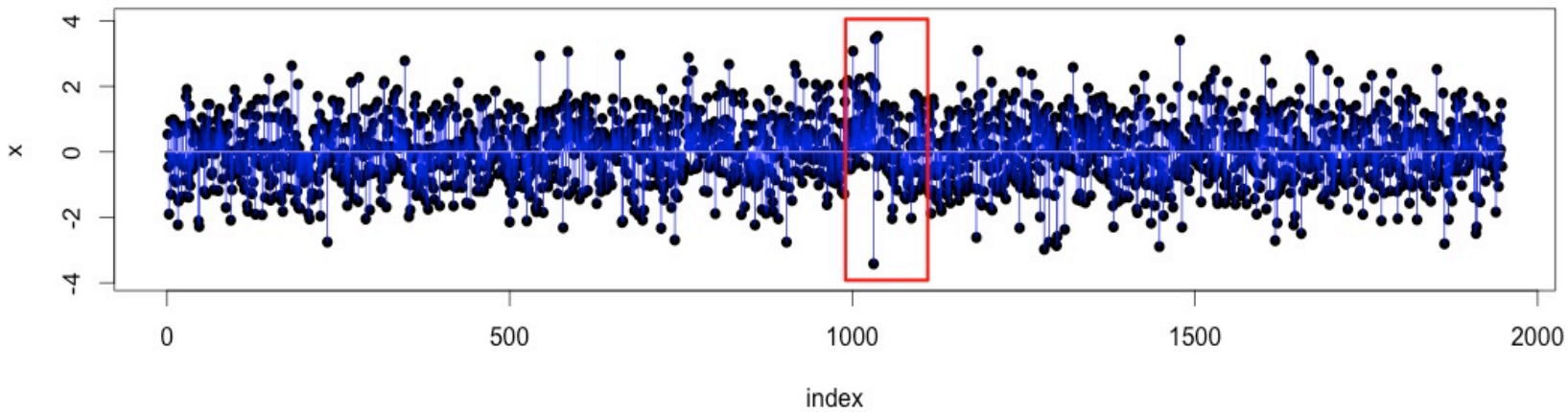
  - Adversarial case – signal is random $\pm A_n$

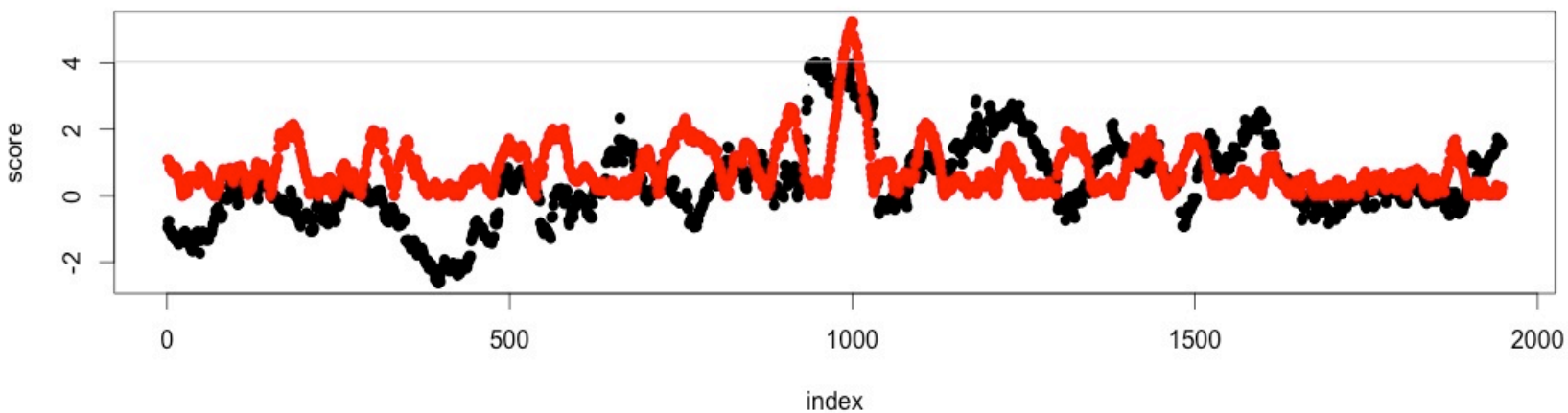
- Comparison with signals of known shape

	Known	Arbitrary
$d \lesssim \log n$	$A_n^2 \gtrsim d^{-1} \log n$	
$d \gtrsim \log n$	$A_n^2 \gtrsim d^{-1} \log n$	$A_n^2 \gtrsim (d^{-1} \log n)^{1/2}$

  - Effect of multiplicity – only $\sqrt{\log n}$ when $d \gtrsim \log n$

Linear vs Quadratic Scan





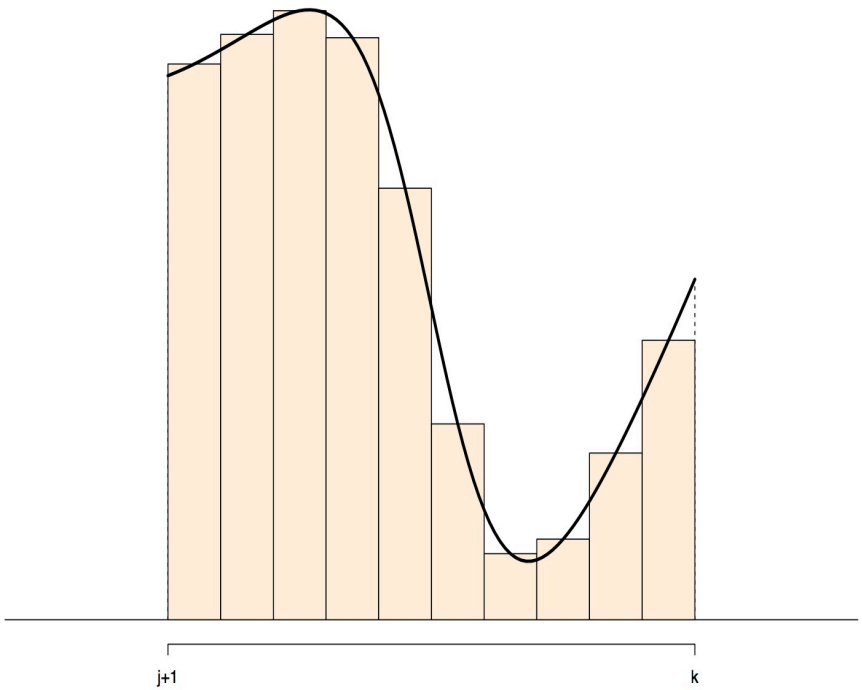
Smooth Signals

- f_n is α times differentiable in that

$$|f^{(\lfloor\alpha\rfloor)}(x) - f^{(\lfloor\alpha\rfloor)}(x')| \leq L|x - x'|^{\alpha - \lfloor\alpha\rfloor}$$

- Optimal rate of detection
  - Determining factors – amplitude, duration and degree of smoothness
- How to scan for smooth signal – combining linear and quadratic statistics
  - Linear statistics most powerful if signal is almost constant
  - Quadratic statistics most powerful if signal changes rapidly

Scan for Smooth Signals



- Divide the segment $(a_n, b_n]$ into l bins, each of size $m = d_n/l$
- Average within each bin

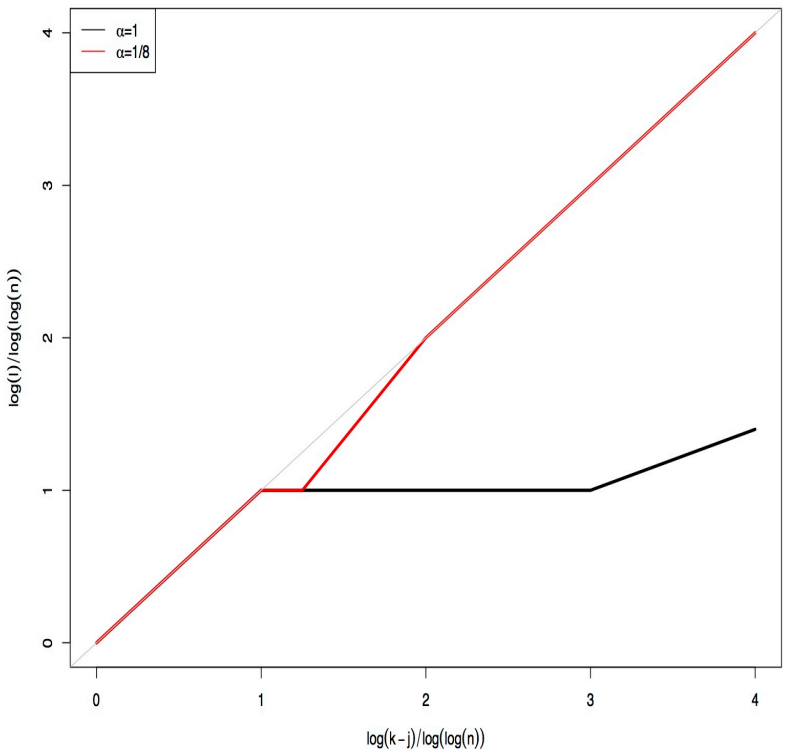
$$Y_j := m^{-1/2} \sum_{i=1}^{m} X_{a_n+(j-1)m+i}$$

- Gather evidence through

$$(l^{1/2} + \log^{1/2} n)^{-1} \sum_{j=1}^{l} (Y_j^2 - 1)$$

- Scan over all putative intervals – $L_n = \max_{a_n, d_n} L_{(a_n, b_n]}$

How many bins?



- if $\alpha \geq 1/4$
  - If $d_n \leq \log n$

    $$l = d_n$$

  - If $\log n < d_n \leq (\log n)^{2\alpha+1}$

    $$l = \log n$$

  - If $d_n > (\log n)^{2\alpha+1}$

    $$l = d_n^{\frac{2}{4\alpha+1}} (\log n)^{-\frac{1}{4\alpha+1}}$$

- if $\alpha < 1/4$ – an extra change
  - If $d_n > (\log n)^{\frac{1}{1-4\alpha}}$

    $$l = d_n$$

Optimal Rate of Detection

With critical value $c_n = c_0\sqrt{\log n}$, detect α times differentiable signals such that

- if $\alpha \geq 1/4$

$$A_n^2 \gtrsim \begin{cases} d_n^{-1}\log n & \text{if } d_n = O((\log n)^{2\alpha+1}) \\ d_n^{-\frac{4\alpha}{4\alpha+1}}(\log n)^{\frac{2\alpha}{4\alpha+1}} & \text{if } d_n \gg (\log n)^{2\alpha+1} \end{cases}$$

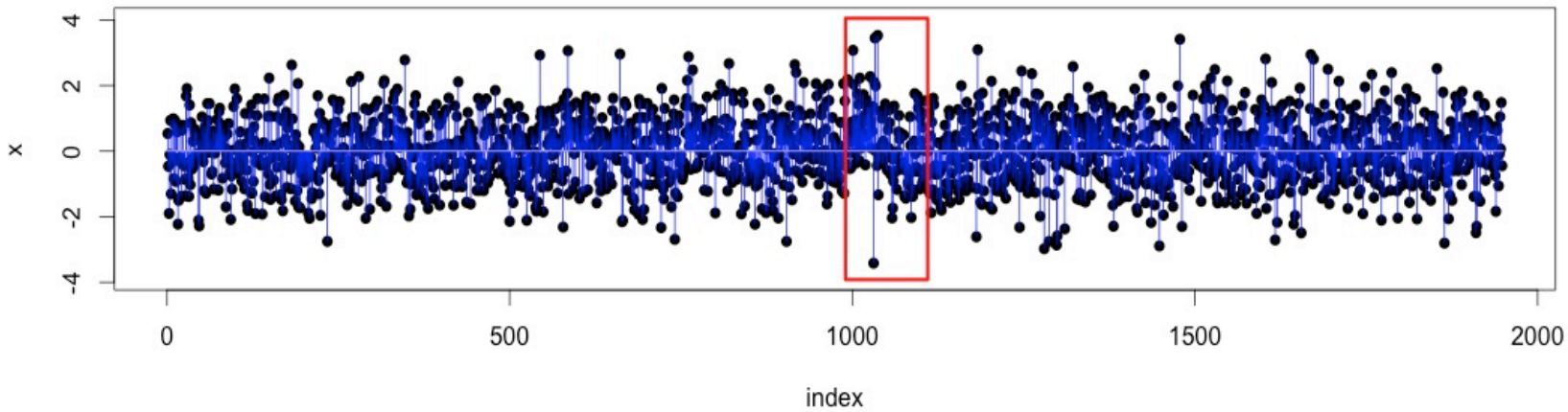
- if $\alpha < 1/4$

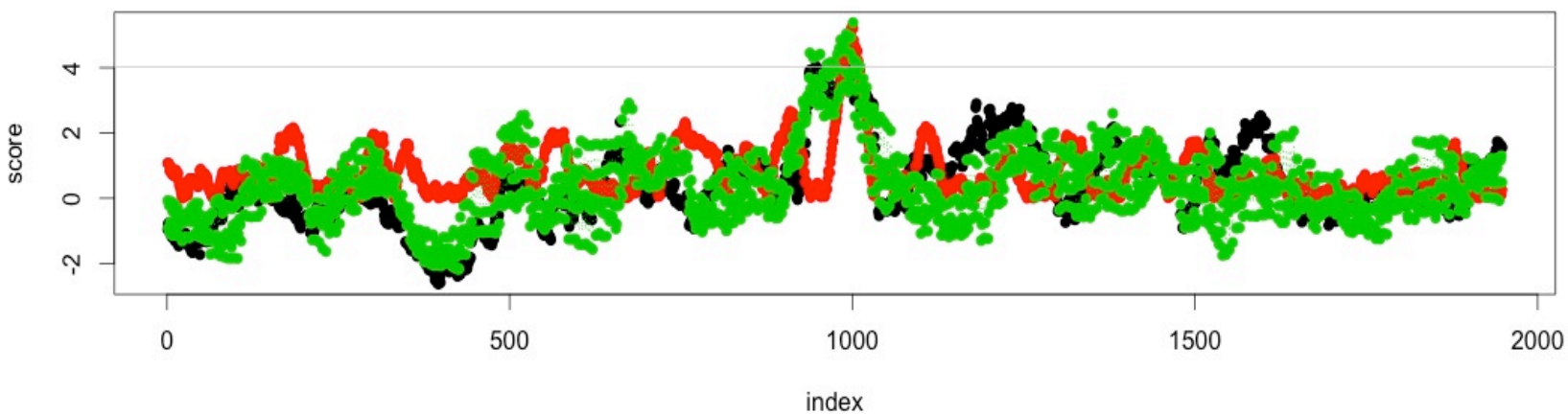
$$A_n^2 \gtrsim \begin{cases} d_n^{-1}\log n & \text{if } d_n = O((\log n)^{2\alpha+1}) \\ d_n^{-\frac{4\alpha}{4\alpha+1}}(\log n)^{\frac{2\alpha}{4\alpha+1}} & \text{if } (\log n)^{2\alpha+1} \ll d_n \ll (\log n)^{1/(1-4\alpha)} \\ (d_n^{-1}\log n)^{1/2} & \text{if } d_n \gg (\log n)^{1/(1-4\alpha)} \end{cases}$$

Compared with the case when the location of signal is known in advance (Ingster)

$$A_n^2 \gtrsim d_n^{-\frac{4\alpha}{4\alpha+1}}$$

Effect of "Binning"





Summary

- Detection of sparse signal segments is a common problem in many high dimensional problems
- Detectability of sparse segments depends on the signal strength, duration, and shape
- Rate optimal detection can be achieved with scan statistics
  - Scan with linear statistics for signal of known shape
  - Scan with quadratic statistics for signal of arbitrary shape
  - Combing linear and quadratic statistics for signals of smoothness
- Ongoing work
  - Adaptation
  - Beyond normality